Wolfgang Ahrens
Iris Pigeot
Editors

Handbook of Epidemiology

Second Edition

SpringerReference

Handbook of Epidemiology

Wolfgang Ahrens • Iris Pigeot
Editors

Handbook of Epidemiology

Second Edition

With 280 Figures and 200 Tables

Springer Reference

Editors

Wolfgang Ahrens
Department of Epidemiological Methods
and Etiologic Research
Leibniz Institute for Prevention Research
and Epidemiology – BIPS
Bremen, Germany

Iris Pigeot
Department of Biometry and Data
Management
Leibniz Institute for Prevention Research
and Epidemiology – BIPS
Bremen, Germany

ISBN 978-0-387-09833-3 ISBN 978-0-387-09834-0 (eBook)
ISBN Bundle 978-0-387-09835-7 (print and electronic bundle)
DOI 10.1007/978-0-387-09834-0
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013946541

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword to the Second Edition

Since 2005, when the first edition of the *Handbook of Epidemiology* was published, the English version of *Wikipedia* that was launched in 2001 has expanded from half a million articles and 1 gigabyte of text to about four million articles and 8 GB of text. A large number of articles deal with scientific and technical topics, including epidemiology and biostatistics. Has a stage been attained when the laborious task of compiling and editing books, particularly of an encyclopedic character, should be considered as not worth the effort, being outpaced by the swift mechanism of writing and updating texts online by volunteer authors?

Probably that stage will never be reached as different media fulfill different roles. Today there is room and need for handbooks occupying the territory between textbooks, with specific contents, perspectives, and targeted readership and all-encompassing encyclopedias of classical or “wiki” type in which the authors have freedom of approach provided they keep a presentation level allowing accessibility to a wide public. The *Handbook of Epidemiology* felicitously combines traits of both types of publications: it has the coverage expected from an up-to-date textbook of epidemiology and shows a variety of approaches by the different authors, who all share a style of presentation making each chapter well readable by epidemiologists and epidemiologists in training. More than a hundred scientifically distinguished authors have contributed to the book which benefits of their cumulated experience of at least five million person-hours of research and teaching. In this second edition, the total number of chapters has been substantially increased from 40 to 60. The sections of the book have been partly reorganized and their new headings neatly indicate the comprehensive coverage of the book: Concepts and Designs in Epidemiology; Methodological Approaches in Epidemiology; Statistical Methods in Epidemiology; Exposure-Oriented Epidemiology; Outcome-Oriented Epidemiology. The handbook is both a reference and a tool for several daily tasks: when writing a paper or reviewing articles and reports it helps checking the appropriateness and the implications of concepts and words, particularly if used in a non-standard way; when teaching and lecturing it may assist in preparing notes and visual aids. In addition going through individual chapters may provide a fresh look even at topics one is already familiar with (or thinks to be so).

As the name indicates, a handbook is designed to be “at hand,” on paper or electronically: I wish many in the epidemiological circles will have this *Handbook of Epidemiology* ready at hand for the consultation it fully deserves.

Florence, Italy Rodolfo Saracci

Foreword to the First Edition

When I was learning epidemiology nearly 50 years ago, there was barely one suitable textbook and a handful of specialized monographs to guide me. Information and ideas in journals were pretty sparse too. That all began to change about 25 years ago and soon we had a plethora of books to consider when deciding on something to recommend to students at every level from beginners to advanced postgraduates. This one is different from all the others. There has never been a single source of detailed descriptive accounts and informed discussions of all the essential aspects of practical epidemiology, written by experts and intended as a desk reference for mature epidemiologists who are in practice, probably already specializing in a particular field, but in need of current information and ideas about every aspect of the state of the art and science. Without a work like this, it is difficult to stay abreast of the times. A comprehensive current overview like this where each chapter is written by acknowledged experts chosen from a rich international pool of talent and expertise makes the task considerably easier.

It had been a rare privilege to receive and read the chapters as they have been written and sent to me through cyberspace. Each added to my enthusiasm for the project. I know and have a high regard for the authors of many of the chapters, and reading the chapters by those I did not know has given me a high regard for them too. The book has a logical framework and structure, proceeding from sections on concepts and methods and statistical methods to applications and fields of current research. I have learned a great deal from all of it, and furthermore I have enjoyed reading these accounts. I am confident that many others will do so too.

John M. Last
Emeritus Professor of Epidemiology
University of Ottawa, Canada

Preface to the Second Edition

The objective of this handbook is to provide a comprehensive overview of the field of epidemiology, bridging the gap between standard textbooks of epidemiology and publications for specialists with a narrow focus on specific areas. The handbook reviews the methodological approaches, statistical concepts, and important subject-matter topics pertinent to the field for which the reader seeks a detailed overview. It thus serves both as a first orientation for the interested reader and a starting point for an in-depth study of a specific area.

The handbook is intended as a reference source and a summarizing overview for professionals involved in health research, health reporting, health promotion, and health system administration. As a thorough guide through the major fields of epidemiology, it also provides a comprehensive resource for public health researchers, physicians, biostatisticians, epidemiologists, and executives in health services.

The first edition of this handbook, which appeared in 2005, was sold out much faster than expected. It had to be reprinted twice and received a subject index, which was missing in the first print. At that time we had already planned to work on a second, improved and expanded edition of the handbook, and we asked all authors for suggestions of further relevant topics to be covered by the handbook. The revision of the existing chapters started with an anonymous peer review of each chapter by authors of other chapters. It turned out that authors reviewed the chapters quite thoroughly making constructive suggestions for improvements and clarifications at the same time. This led to a careful revision of each chapter, several of which have undergone a fundamental renewal and update. We are very grateful to all authors for engaging in this fruitful process which helped to substantially improve the second edition. The second edition now covers several new topics in its 60 chapters, 20 more than the first edition, where topics are addressed that are usually only marginally represented in the specific literature or in standard textbooks, such as planning of epidemiological studies, data management, ethical aspects, practical fieldwork, epidemiology in developing countries, quality control, and good epidemiological practice.

The content of the handbook is arranged according to a new structure. It now contains five major parts ranging from concepts and designs (Part I), methodological approaches (Part II), statistical methods typically applied in epidemiological studies (Part III), various dimensions of disease determinants, i.e., exposures (Part IV), and major disease outcomes (Part V), where "exposure-oriented" and "outcome-oriented" epidemiology represent the two sides of the same coin. If the research question emphasizes disease determinants, the corresponding studies are usually classified as exposure-oriented. If, in contrast, a disease or another health-related event is the focus, we speak of "outcome-oriented" studies, in which risk factors for the specific disease are searched for. The above distinction is nevertheless somewhat arbitrary, as some research areas are not easily assigned to either category, e.g., molecular epidemiology, screening, clinical epidemiology, and infectious disease epidemiology.

In the following, we give an overview of the various chapters in the five parts, where every chapter of the handbook follows a similar structure. Each chapter stands on its own, giving an overview of the topic and the most important problems and approaches, which are supported by examples, practical applications, and illustrations. It serves as an introduction that allows one to enter a new field by addressing basic concepts and standard approaches, as well as recent advances and new perspectives. It thus also conveys more advanced knowledge for the reader who is familiar with the given topic by reflecting the state of the art and future prospects. Each chapter provides references both for the novice as well as for the advanced reader. Of course, some basic understanding of the concepts of probability, sampling distribution, estimation, and hypothesis testing will enable the reader to benefit from the statistical concepts primarily presented in Part III, and from the comprehensive discussion of empirical methods in the other parts.

After introducing the historical development of epidemiology at the beginning of **Part I**, the epistemological background and the conceptual building blocks of epidemiology such as models for causation and statistical ideas are described. The latter aspect is deepened with an overview of the various risk measures used in epidemiological studies. These measures depend on the type of data and the study design chosen to answer the research question: Descriptive studies provide the basic information for health reporting. Observational studies like cohort studies, case-control studies, and some modern study designs as well as experimental studies like intervention trials and cluster randomized trials, serve to examine associations and hypothesized causal relationships. The chapter on community-based health promotion builds the bridge between knowledge about the etiology of diseases and its application for prevention. Part I concludes with the implementation of classical epidemiological study designs using innovative internet-based approaches.

Part II of the handbook addresses the major methodological challenges faced by epidemiologists. These concern practical problems to be handled when conducting an epidemiological study such as key aspects of the planning of studies in general, their quality control, primary data collection, and exposure assessment. Further topics are major methodological issues that typically arise in observational studies such as the problem of misclassification, the two concepts of interaction and confounding as well as the problem of how to perform bias and sensitivity analyses. These more theoretical approaches are complemented by describing epidemiological methods for disease monitoring, surveillance and screening, where the latter aims at early detection of chronic diseases. Particular focus is on the problems related to health systems in developing countries and the resulting demands for epidemiological research as well as on the impact of the results of epidemiological studies on political decisions and the health system in developed countries. Part II closes with a discussion of human rights and responsibilities that have to be considered at the different stages of an epidemiological study.

The arsenal of statistical concepts and methods to be found in epidemiology is covered by **Part III**. It starts with a new general introduction into the basic statistical concepts that help to understand statistical thinking, where three complementing approaches to statistical inference are distinguished: the frequentist, the likelihood, and the Bayesian approach. Bayesian methods that try to integrate the evidence from the data with knowledge of the underlying scientific theory are introduced in a new chapter with emphasis on their application in epidemiological studies. The necessary steps during data management which are the prerequisites for any statistical analysis are now also described in the handbook in addition to another prerequisite, i.e., the sample size determination. Advanced techniques, such as fractional polynomials, are introduced for dose-effect analysis where exposures and/or outcomes are described by continuous variables. Since the relation between exposures and outcomes, which is the essence of epidemiology, is mostly represented by regression models, it is not surprising that the corresponding chapter is the longest in the handbook. Models where the outcome variables are in the form of a waiting time until a specific event, e.g., death, are discussed separately. Given that in practice data are often erroneous or missing, methods to handle the ensuing problems are presented in two chapters. The special problem of correlated data is accounted for in a new chapter on generalized estimating equations. Further issues in this statistical part of the handbook concern meta-analysis: the art of drawing joint conclusions from the results of several studies in order to put these conclusions on firmer ground; the analysis of spatial data where the values of the principal explanatory variable are geographical locations; and the analysis of genetic data, especially their high dimensionality, which requires specific methods for their analyses and directed acyclic graphs that play a major role under two different perspectives, i.e., to investigate the complex interplay in high-dimensional datasets and to capture causal interrelationships between determinants and potential outcomes.

Although each epidemiological study entails its own peculiarities and specific problems related to its design and realization, depending on the field of application, common features can be identified. Many important, partly classical, partly recent, applications of epidemiology of general interest to public health are defined by specific exposures. This exposure-oriented view on methods and design is taken in **Part IV**, which starts with the introduction of methods especially devoted to life course epidemiology. This new chapter is followed by the presentation of the main exposure-oriented fields such as social, occupational, environmental, nutritional, and reproductive epidemiology, but also more recent applications, such as the inclusion of molecular markers, are described. Clinical epidemiology and pharmacoepidemiology are growing areas where knowledge about the interplay between medical treatments and lifestyle factors and their joint effect on health outcomes is generated. Another exposure that is of special relevance with respect to the prevention of diseases is the individual's physical activity, which is discussed not only with respect to its public health relevance but also the current methodology to measure this exposure. A classical exposure with continuing interest to epidemiologists is ionizing radiation, where medical treatments add to occupational and environmental exposures. The specific methodological approaches to quantify these exposures are described in another new chapter.

Part V, the final part of this handbook, takes a disease-oriented perspective. It addresses specific methodological challenges to investigate prominent health outcomes and describes their epidemiology. The list of possible endpoints is almost endless. We selected ten disease groups because of their high public health relevance and/or their specific methodological challenges. These are infectious diseases, cardiovascular diseases, cancer, musculoskeletal disorders, and – newly included – overweight and obesity, asthma and allergies, dental diseases, digestive diseases, psychiatric disorders, and, last but not least, diabetes.

The editors are indebted to the knowledgeable experts for their valuable contributions, their enthusiastic support in producing this handbook, and their patience during the writing of the second edition, which took longer than we expected. We would like to thank Frauke Günther, Marc Suling, Claudia Börnhorst, and Angela Kammrad for their technical support. We are particularly grateful for the continuous and outstanding engagement of Regine Albrecht. Without her support, her patience with us and the contributors, and her remarkable autonomous management of the editorial work, this volume would not have been possible.

Islamorada, USA
January 2013

Wolfgang Ahrens
Iris Pigeot

Preface to the First Edition

The objective of this book is to provide a comprehensive overview of the field of epidemiology, bridging the gap between standard textbooks of epidemiology and publications for specialists with a narrow focus on specific areas. It reviews the key issues, methodological approaches and statistical concepts pertinent to the field for which the reader seeks a detailed overview. It thus serves both as a first orientation for the interested reader and a starting point for an in-depth study of a specific area, as well as a quick reference and a summarizing overview for the expert.

The handbook is intended as a reference source for professionals involved in health research, health reporting, health promotion, and health system administration and related experts. It covers the major aspects of epidemiology and may be consulted as a thorough guide for specific topics. It is therefore of interest for public health researchers, physicians, biostatisticians, epidemiologists, and executives in health services.

The broad scope of the book is reflected by four major parts that facilitate an integration of epidemiological concepts and methods, statistical tools, applications, and epidemiological practice. The various facets are presented in 39 chapters and a general introduction to epidemiology. The latter provides the framework in which all other chapters are embedded and gives an overall picture of the whole handbook. It also highlights specific aspects and reveals the interwoven nature of the various research fields and disciplines related to epidemiology. The book covers topics that are usually missing from standard textbooks and that are only marginally represented in the specific literature, such as ethical aspects, practical fieldwork, health services research, epidemiology in developing countries, quality control, and good epidemiological practice. It also covers innovative areas, e.g., molecular and genetic epidemiology, modern study designs, and recent methodological developments.

Each chapter of the handbook serves as an introduction that allows one to enter a new field by addressing basic concepts, but without being too elementary. It also conveys more advanced knowledge and may thus be used as a reference source for the reader who is familiar with the given topic by reflecting the state of the art and future prospects. Of course, some basic understanding of the concepts of probability, sampling distribution, estimation, and hypothesis testing will help the reader to profit from the statistical concepts primarily presented in Part II and from the comprehensive discussion of empirical methods in the other parts. Each chapter is intended to stand on its own, giving an overview of the topic and the most important problems and approaches, which are supported by examples, practical applications, and illustrations. The basic concepts and knowledge, standard procedures and methods are presented, as well as recent advances and new perspectives. The handbook provides references both to introductory texts and to publications for the advanced reader.

The editors dedicate this handbook to Professor Eberhard Greiser, one of the pioneers of epidemiology in Germany. He is the founder of the Bremen Institute for Prevention Research and Social Medicine (BIPS), which is devoted to research into the causes and the prevention of disease. This institute, which started as a small enterprise dedicated to cardiovascular prevention, has grown to become one of the most highly regarded research institutes for epidemiology and public health in Germany. For almost 25 years Eberhard Greiser has been a leader in the field of epidemiology, committing his professional career to a critical appraisal of health practices for the benefit of us all. His major interests have been in pharmaceutical care and social medicine. In recognition of his contributions as a researcher and as a policy advisor to the advancement of the evolving field of epidemiology and public health in Germany we take his 65th birthday in November 2003 as an opportunity to acknowledge his efforts by editing this handbook.

The editors are indebted to knowledgeable experts for their valuable contributions and their enthusiastic support in producing this handbook. We thank all the colleagues who critically reviewed the chapters: Klaus Giersiepen, Cornelia Heitmann, Katrin Janhsen, Jürgen Kübler, Hermann Pohlabeln, Walter Schill, Jürgen Timm, and especially Klaus Krickeberg for his never-ending efforts. We also thank Heidi Asendorf, Thomas Behrens, Claudia Brünings-Kuppe, Andrea Eberle, Ronja Foraita, Andrea Gottlieb, Frauke Günther, Carola Lehmann, Anette Lübke, Ines Pelz, Jenny Peplies, Ursel Prote, Achim Reineke, Anke Suderburg, Nina Wawro, and Astrid Zierer for their technical support. Without the continuous and outstanding engagement of Regine Albrecht - her patience with us and the contributors and her remarkable autonomy - this volume would not have been possible. She has devoted many hours to our handbook over and above her other responsibilities as administrative assistant of the BIPS. Last but not least we are deeply grateful to Clemens Heine of Springer for his initiative, support, and advice in realizing this project and for his confidence in us.

June 2004

Wolfgang Ahrens
Iris Pigeot
Bremen

About the Editors

Wolfgang Ahrens Department of Epidemiological Methods and Etiologic Research, Leibniz Institute for Prevention Research and Epidemiology – BIPS Bremen, Germany

Wolfgang Ahrens is Professor for Epidemiological Methods at the Department of Mathematics and Computer Science of the University of Bremen since 2003. He is deputy director of the Leibniz Institute for Prevention Research and Epidemiology – BIPS, and head of the Department of Epidemiological Methods and Etiologic Research. Wolfgang Ahrens is an epidemiologist with considerable experience in leading population-based (inter)national multi-center studies involving primary data collection. His research focuses on epidemiological methods with particular emphasis on exposure assessment and the etiology and primary prevention of non-communicable diseases, especially cancer and nutrition- and lifestyle-related disorders. He coordinates two large European cohort studies of more than 16,000 children from eight countries, IDEFICS and I.Family, and he is member of the board of scientific directors of the German National Cohort including 200,000 German residents. He has published more than 220 international peer-reviewed papers and 16 book chapters. He has (co)authored three books, the most recent being a German textbook “Epidemiological Methods” (Springer 2012) and edited six books.

Iris Pigeot Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology – BIPS Bremen, Germany

Iris Pigeot is Professor of Statistics focused on Biometry and Methods in Epidemiology at the Department of Mathematics and Computer Science of the University of Bremen since 2001. She is Director of the Leibniz Institute for Prevention Research and Epidemiology – BIPS, and head of the Department of Biometry and Data Management. Her research activities focus on bioequivalence studies, graphical models, and genetic epidemiology. In recent years, she has widened the spectrum of her research to include the use of secondary data in the research of pharmaceutical drug safety as well as primary prevention and its evaluation, especially for childhood obesity. She received several teaching awards: the "Medal for Excellent Teaching" at the University of Dortmund in 1994, the "Award for Quality Teaching" at the University of Munich in 1996, and the "Berninghausen Award for Excellent Teaching and its Innovation" at the University of Bremen in 2008. In 2010, the IBS-DR awarded her the Susanne-Dahms medal for special accomplishments in the field of biometry.

Iris Pigeot and Wolfgang Ahrens are both editors of the book series "Epidemiology & Public Health" published by Springer (Heidelberg)

Contents

Volume 1

Volume 2

Volume 3

Volume 4

Volume 5

Contributors

Wolfgang Ahrens Department of Epidemiological Methods and Etiologic Research, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

Faculty of Mathematics/Computer Science, University of Bremen, Bremen, Germany

Karin Bammann Institute for Public Health and Nursing Research, Faculty of Human and Health Sciences, University of Bremen, Bremen, Germany

Olga Basso Department of Obstetrics and Gynecology and Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC, Canada

Heiko Becher Unit of Epidemiology and Biostatistics, Ruprecht Karls-University Heidelberg, Institute of Public Health, Heidelberg, Germany

Jacques Benichou Department of Biostatistics, University of Rouen Medical School and Rouen University Hospital, Rouen Cedex, France

Yoav Ben-Shlomo School of Social and Community Medicine, University of Bristol, Bristol, UK

Marianne Berwick Division of Epidemiology and Biostatistics, University of New Mexico, Albuquerque, NM, USA

Caroline Beunckens BELGACOM, Brussels, Belgium

Heike Bickeböller Department of Genetic Epidemiology, University Medical School, Georg-August-University of Göttingen, Göttingen, Germany

John F. Bithell St Peter's College, University of Oxford, Oxford, UK

Maria Blettner Institute for Medical Biostatistics, Epidemiology and Informatics, Johannes Gutenberg University, Mainz, Germany

Heiner Boeing Department of Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany

Paolo Boffetta Institute for Translational Epidemiology and Tisch Cancer Institute, Mount Sinai School of Medicine, New York, NY, USA

International Prevention Research Institute, Lyon, France

Freddie Bray Section of Cancer Information, International Agency for Research on Cancer, Lyon, France

Adam R. Brentnall Wolfson Institute of Preventive Medicine, Queen Mary University of London, London, UK

Norman E. Breslow Department of Biostatistics, University of Washington, Seattle, WA, USA

James W. Buehler Public Health Surveillance Program Office, Office of Surveillance, Epidemiology, and Laboratory Services, Centers for Disease Control and Prevention, Atlanta, GA, USA

Stephen L. Buka Department of Epidemiology, Brown University, Providence, RI, USA

Reinhard Busse Department of Health Care Management, Berlin University of Technology, Berlin, Germany

Jeffrey S. Buzas Department of Mathematics and Statistics, University of Vermont, Burlington, VT, USA

Michael J. Campbell Medical Statistics Group, Health Services Research, ScHARR, University of Sheffield, Sheffield, UK

Asit K. Chakraborty BIKALPA, Koramangala, Bangalore, India

Tarani Chandola CCSR and Social Statistics, University of Manchester, Manchester, UK

David I. Conway University of Glasgow Dental School, Glasgow, UK

Susan T. Cookson CAPT, US Public Health Service, Center for Global Health, Centers for Disease Control and Prevention, Atlanta, GA, USA

Sylvaine Cordier INSERM U1085 (National Institute of Health and Medical Research), University of Rennes I, Rennes, France

Johannes J. M. van Delden Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

Jeroen Douwes Centre for Public Health Research, School of Public Health, Massey University Wellington, Wellington, New Zealand

Janet D. Elashoff Statistical Solutions, Saugus, MA, USA

Esther Erdei Division of Epidemiology and Biostatistics, University of New Mexico, Albuquerque, NM, USA

John W. Farquhar Stanford Prevention Research Center, Stanford University School of Medicine, Stanford, CA, USA

Arlène Fink Public Health - Health Services/Med-GIM, Palisades, CA, USA

Katherine M. Flegal Office of the Director, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, USA

Ronja Foraita Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

Silvia Franceschi Infections and Cancer Epidemiology Group, International Agency for Research on Cancer, Lyon, France

Edeltraut Garbe Department of Clinical Epidemiology, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

Faculty of Human and Health Sciences, University of Bremen, Bremen, Germany

Christian A. Gericke The Wesley Research Institute and University of Queensland, School of Population Health, Brisbane, Australia

Mika Gissler Information Department, THL - National Institute for Health and Welfare, Helsinki, Finland

David C. Goff Jr. Department of Epidemiology, Colorado School of Public Health, Aurora, CO, USA

Sander Greenland Department of Epidemiology, School of Public Health, University of California, Los Angeles, CA, USA

Department of Statistics, UCLA College of Letters and Science, Los Angeles, CA, USA

Paul Gustafson Department of Statistics, University of British Columbia, Vancouver, BC, Canada

Gordon H. Guyatt Faculty of Health Sciences, Departments of Clinical Epidemiology and Biostatistics and of Medicine, McMaster University, Hamilton, ON, Canada

Timo Hakulinen Finnish Cancer Registry - Institute for Statistical and Epidemiological Cancer Research, Helsinki, Finland

Leonhard Held Division of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich, Zurich, Switzerland

Frank B. Hu Departments of Nutrition and Epidemiology, Harvard School of Public Health, Boston, MA, USA

Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

Ivy Jansen Research Institute for Nature and Forest (INBO), Brussels, Belgium

Anita Kar School of Health Sciences, University of Pune, Pune, MH, India

Philip H. Kass Department of Population Health and Reproduction, University of California School of Veterinary Medicine, Davis, CA, USA

Michael G. Kenward Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK

Brian K. Kit Division of Health and Nutrition Examination Surveys, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, USA

Epidemic Intelligence Service, Scientific Education and Professional Development Program Office, Centers for Disease Control and Prevention, Hyattsville, MD, USA

Ulrike Krahn Institute for Medical Biostatistics, Epidemiology and Informatics, Johannes Gutenberg University, Mainz, Germany

Lothar Kreienbrock Department for Biometry, Epidemiology and Information Processing, University of Veterinary Medicine Hannover, Hannover, Germany

Mirjam Kretzschmar Julius Centre for Health Sciences and Primary Care University Medical Centre Utrecht, CX Utrecht, The Netherlands

Centre for Infectious Disease Control, RIVM, MA Bilthoven, The Netherlands

Klaus Krickeberg University of Paris V (retired), Bielefeld, Germany

Meena Kumari Department of Epidemiology and Public Health, University College London, London, UK

Diana Kuh MRC Unit for Lifelong Health and Ageing, University College London, London, UK

Darwin R. Labarthe Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

Michael F. Leitzmann Department of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany

Stanley Lemeshow Dean, College of Public Health, The Ohio State University, Columbus, OH, USA

Hubert G. Leufkens Utrecht Institute for Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, The Netherlands

Lorna M. D. Macpherson University of Glasgow Dental School, Glasgow, UK

Barrie M. Margetts Faculty of Medicine, Public Health Nutrition, University of Southampton, Southampton, UK

Michael Marmot Department of Epidemiology and Public Health, University College London, London, UK

Giuseppe Matullo Genomic Variation in Human Population and Complex Diseases, Human Genetics Foundation (HuGeF), Turin, Italy

Department of Medical Science, University of Turin, Turin, Italy

Alex D. McMahon University of Glasgow Dental School, Glasgow, UK

Franco Merletti Unit of Cancer Epidemiology and Centre for Oncologic Prevention, Department of Medical Sciences, University of Turin, Turin, Italy

Hiltrud Merzenich Institute for Medical Biostatistics, Epidemiology and Informatics, University Medical Center, Johannes Gutenberg University Mainz, Mainz, Germany

Anthony B. Miller Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

Dario Mirabelli Unit of Cancer Epidemiology and Centre for Oncologic Prevention, Department of Medical Sciences, University of Turin, Turin, Italy

Gita Mishra School of Population Health, University of Queensland, Brisbane, Australia

Geert Molenberghs Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), University of Hasselt and Catholic University Leuven (KU Leuven), Diepenbeek, Belgium

Alfredo Morabia Center for the Biology of Natural Systems, Queens College City University of New York, New York, Flushing, NY, USA

Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA

Gila Neta Division of Cancer Control and Population Sciences and Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA

Cynthia L. Ogden Division of Health and Nutrition Examination Surveys, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, USA

Jørn Olsen Department of Public Health, Section for Epidemiology, Aarhus University, Aarhus C, Denmark

Department of Epidemiology, Center for Health Sciences (CHS), UCLA School of Public Health, Los Angeles, CA, USA

Mari Palta Department of Population Health Sciences and Department of Biostatistics and Medical Informatics, University of Wisconsin School of Medicine and Public Health, Wisconsin, WI, USA

Daniela Paolotti Computational Epidemiology Lab, ISI Foundation, Turin, Italy

D. Maxwell Parkin Clinical Trials Service Unit & Epidemiological Studies Unit, Department of Medicine, University of Oxford, Oxford, UK

Neil Pearce Department of Medical Statistics, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

Centre for Public Health Research, School of Public Health, Massey University Wellington, Wellington, New Zealand

Iris Pigeot Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

Faculty of Mathematics/Computer Science, University of Bremen, Bremen, Germany

Costanza Pizzi Cancer Epidemiology Unit, Department of Medical Sciences, CPO-Piemonte and University of Turin, Turin, Italy

Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK

Martyn Plummer Infections and Cancer Epidemiology Group, International Agency for Research on Cancer, Lyon, France

Hermann Pohlabeln Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiolgy - BIPS, Bremen, Germany

Preetha Rajaraman South Asia Region Center for Global Health, US National Cancer Institute, New Delhi, India

Raoult Ratard Louisiana Department of Health and Hospitals, Office of Public Health, Infectious Disease Epidemiology Section, New Orleans, LA, USA

Achim Reineke Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiolgy - BIPS, Bremen, Germany

Lorenzo Richiardi Unit of Cancer Epidemiology and Centre for Oncologic Prevention, Department of Medical Sciences, University of Turin, Turin, Italy

Hilkka Riihimäki Centre of Expertise for Health and Work Ability (retired), Finnish Institute of Occupational Health (FIOH), Espoo, Finland

Douglas Robertson University of Glasgow Dental School, Glasgow, UK

Kenneth J. Rothman RTI Health Solutions, Research Triangle Institute, Research Triangle Park, NC, USA

Måns Rosén SBU – The Swedish Council on Technology, Assessment in Health Care, Stockholm, Sweden

Jonathan M. Samet Director, University of Southern California (USC) Institute for Global Health, Professor and Flora L. Thornton Chair, Department of Preventive Medicine Keck School of Medicine of USC

Peter D. Sasieni Wolfson Institute of Preventive Medicine, Queen Mary University of London, London, UK

Thomas Schäfer Department of Economics and Information Technology, University of Applied Sciences Gelsenkirchen, Bocholt, Germany

Walter Schill Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiolgy - BIPS, Bremen, Germany

Peter Schlattmann Institute of Medical Statistics, Computer Sciences and Documentation, Jena University Hospital, Jena, Germany

Daniela Schmid Department of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany

Holger J. Schünemann Faculty of Health Sciences, Departments of Clinical Epidemiology and Biostatistics and of Medicine, McMaster University, Hamilton, ON, Canada

Matthias B. Schulze Department of Molecular Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany

Jacob Spallek Department of Epidemiology & International Public Health, University of Bielefeld – School of Public Health, Bielefeld, Germany

Leonard A. Stefanski Department of Statistics, North Carolina State University, Carolina, NC, USA

Patricia A. Stewart Stewart Exposure Assessments, LLC, Arlington, VA, USA

Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD, USA (retired)

Susanne Straif-Bourgeois Louisiana Department of Health and Hospitals, Office of Public Health, Infectious Disease Epidemiology Section, New Orleans, LA, USA

Samy Suissa Division of Clinical Epidemiology, McGill University Health Center, Montreal, QC, Canada

Reijo Sund Service Systems Research Unit, THL - National Institute for Health and Welfare, Helsinki, Finland

Ezra S. Susser Mailman School of Public Health, Columbia University and New York State Psychiatric Institute, New York, NY, USA

Lloyd R. Sutherland University of Calgary, Calgary, AB, Canada

Herbert Thijs Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), University of Hasselt and Catholic University Leuven (KU Leuven), Diepenbeek, Belgium

Duncan C. Thomas Department of Preventive Medicine, Division of Biostatistics, University of Southern California, Keck School of Medicine, Los Angeles, CA, USA

Antje Timmer Department of Clinical Epidemiology, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

Salina M. Torres Division of Epidemiology and Biostatistics, University of New Mexico, Albuquerque, NM, USA

Tor D. Tosteson Dartmouth College, Lebanon, NH, USA

Maren Vens Institute of Medical Biometry and Statistics, University of Lübeck, University Medical Center Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

Geert Verbeke Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), University of Hasselt and Catholic University Leuven (KU Leuven), Diepenbeek, Belgium

Paolo Vineis Chair of Environmental Epidemiology, School of Public Health, Imperial College London, London, UK

Emma W. Viscidi Department of Epidemiology, Brown University, Providence, RI, USA

Henryk Wicke Institute for Medical Biostatistics, Epidemiology and Informatics, University Medical Center, Johannes Gutenberg University Mainz, Mainz, Germany

Pascal Wild Direction scientifique (Scientific Management), Occupational Health and Safety Institute (INRS), Vandoeuvre-lès-Nancy, France

Hajo Zeeb Department of Prevention and Evaluation, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

Faculty of Human and Health Sciences, University of Bremen, Bremen, Germany

Andreas Ziegler Institute of Medical Biometry and Statistics, University Medical Center Schleswig-Holstein, Campus Lübeck and Center for Clinical Trials, University of Lübeck, Lübeck, Germany

Part I

Concepts and Designs in Epidemiology

An Introduction to Epidemiology

1

Wolfgang Ahrens, Klaus Krickeberg, and Iris Pigeot

Contents

W. Ahrens (✉)
Department of Epidemiological Methods and Etiologic Research, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

Faculty of Mathematics/Computer Science, University of Bremen, Bremen, Germany

K. Krickeberg
University of Paris V (retired), Bielefeld, Germany

I. Pigeot
Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

Faculty of Mathematics/Computer Science, University of Bremen, Bremen, Germany

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_42,

1.1 Scope

This chapter introduces the key concepts of epidemiology and summarizes their development. It highlights major areas of this discipline that are dealt with in detail in the chapters of this handbook and describes their interrelationship. The whole research area covered by the term "epidemiology" has been systemized in different ways and from various perspectives.

A first distinction differentiates between studies that are designed to evaluate the health effects of planned interventions and studies that are designed to just observe health-related characteristics and events in population groups. The first type of studies is called *experimental* because they introduce a certain measure to improve a health outcome of interest in a defined population. This measure is usually assigned to study subjects by random allocation and controlled by the investigator. Often experimental epidemiology is simply equated with randomized controlled trials. Clinical trials to study the efficacy of a treatment as, for example, a new drug, are a special type within this category. They are to be distinguished from trials of preventive interventions. The second type of studies, typically referred to as *observational*, studies situations as they present themselves without intervening. Such studies may describe the frequency or incidence of health outcomes or health determinants as well as their temporal or geographical variation. They may also be used to describe the natural history of a disease. If they aim to understand the etiology of a disease, they often exploit the so-called natural experiments in that they compare, for example, subjects exposed to a specific risk factor with non-exposed subjects. A typical example is the investigation of the influence of a risk factor like air pollution on a health outcome like asthma.

This gives rise to another, less clearly defined, classification of observational studies, that is, *explanatory* (or *analytical*) vs. *descriptive*. The objective of an explanatory study is to contribute to the search of causes for health-related events, in particular by isolating the effects of specific factors. The causality-oriented perspective distinguishes this type of studies from descriptive studies. In practice this distinction often corresponds to the use of different sources of data: In descriptive epidemiology often data are used that are routinely collected for various reasons, for example, for administrative purposes, whereas in explanatory or analytical epidemiology, data are usually collected for the study purpose itself. The term "descriptive epidemiology" covers the field of health reporting and is closely related to health statistics where reported data are usually stratified by time, place of residence, age, sex, and social position.

Observational epidemiology can also be divided into *exposure-oriented* and *outcome-oriented* epidemiology. These two research areas represent the two sides of the same coin. Insofar this distinction is more systematic rather than substantive. If the research question emphasizes disease determinants, for example, environmental or genetic factors, the corresponding studies usually are classified as exposure-oriented. If, in contrast, a disease or another health-related event like lung cancer or osteoarthritis is the focus, we speak of outcome-oriented studies in which risk factors for the specific disease are searched for.

After having introduced some common ways to classify the epidemiological research areas, we explain the definition of epidemiology, describe its scope, and illustrate its relations to adjacent research areas in Sect. 1.2. Section 1.3 gives a brief overview of the historical development of epidemiology and its applied areas. Section 1.4 provides an overview of the basic concepts related to the epidemiological study designs and describes specific applied research fields. Key concepts of statistical reasoning and major statistical methods are introduced in Sect. 1.5. Section 1.6 discusses "exposure-oriented" and "outcome-oriented" epidemiology and highlights some major research areas in both fields.

1.2 Epidemiology and Related Areas

Various disciplines contribute to the investigation of determinants of human health and disease, to the improvement of health care, and to the prevention of illness. These contributing disciplines stem from four major scientific areas, first from basic biomedical sciences such as biology, physiology, biochemistry, molecular genetics, and pathology; second from clinical sciences such as oncology, gynecology, orthopedics, obstetrics, cardiology, internal medicine, urology, radiology, and pharmacology; third from biostatistics and demography; and fourth from social and behavioral sciences as well as public health with epidemiology as its core.

1.2.1 Definition and Purpose of Epidemiology

One of the first definitions of epidemiology that was most frequently used at that time was given by MacMahon and Pugh (1970):

> Epidemiology is the study of the distribution and determinants of disease frequency in man.

This relatively simplistic definition was further developed later on as described below. The three components of this definition, that is, frequency, distribution, and determinants, embrace the basic principles and approaches in epidemiological research. The measurement of disease *frequency* relates to the quantification of disease occurrence in human populations. Such data are needed for further investigations of patterns of disease in subgroups of the population. This involves "… describing the *distribution* of health status in terms of age, sex, race, geography, etc., …" (MacMahon and Pugh 1970). The methods used to describe the distribution of diseases may be considered as a prerequisite to identify the *determinants* of human health and disease.

The above definition is based on two fundamental assumptions: first, the occurrence of diseases in populations is not a purely random process, and second, it is determined by causal and preventive factors (Hennekens and Buring 1987). As mentioned above, these factors have to be searched for systematically in populations defined by place, time, or otherwise. Different ecological models have been used to describe the interplay of these factors, which relate to host, agent,

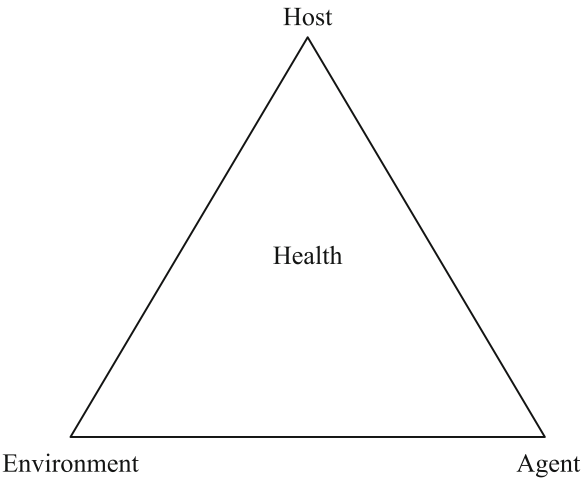


Fig. 1.1 The epidemiological triangle

and environment. Changing any of these three forces, which constitute the so-called epidemiological triangle (Fig. 1.1), will influence the balance among them and thereby increase or decrease the disease frequency (Mausner and Bahn 1974).

Thus, the search for etiological factors in the development of diseases and disorders is one of the main concerns of epidemiology. Complementary to the epidemiological triangle, the triad of time, place, and person is often used by epidemiologists to describe the distribution of diseases and their determinants. Determinants that influence health may consist of behavioral, cultural, social, psychological, biological, or physical factors. The determinants by time may relate to increase/decrease over the years, seasonal variations, or sudden changes of disease occurrence. Determinants by place can be characterized by country, climate zone, residence, and, more general, by geographical region. Personal determinants include age, sex, ethnic group, genetic traits, and individual behavior. Studying the interplay between time, place, and person helps to identify the etiological agent and the environmental factors as well as to describe the natural history of the disease, which then enables the epidemiologist to define targets for intervention with the purpose of disease prevention (Detels 2002). This widened perspective is reflected in a more comprehensive definition of epidemiology as given by Last (2001):

> The study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to control of health problems.

It has been further specified in the subsequent edition of the Dictionary of Epidemiology (Porta 2008):

> The study of the occurrence and distribution of health-related states or events in specified populations, including the study of the determinants influencing such states, and the application of this knowledge to control the health problems.

In this broader sense, health-related states or events include "diseases, causes of death, behaviors, reactions to preventive programs, and provision and use of health services," while determinants refer to "all the physical, biological, social, cultural, economic and behavioral factors that influence health" (Porta 2008). According to this widely accepted definition, the final aim of epidemiology is "to promote, protect, and restore health." Hence, the major goals of epidemiology may be defined from two overlapping perspectives. The first is a biomedical perspective looking primarily at the etiology of diseases and the disease process itself. This includes:

- The description of the disease spectrum, the syndromes of the disease, and the disease entities to learn about the various outcomes that may be caused by particular pathogens
- The description of the natural history, that is, the course of the disease to improve the diagnostic accuracy which is a major issue in clinical epidemiology
- The investigation of physiological or genetic variables in relation to influencing factors and disease outcomes to decide whether they are potential risk factors, disease markers, or indicators of early stages of disease
- The identification of factors that are responsible for the increase or decrease of disease risks in order to obtain the knowledge necessary for primary prevention
- The prediction of disease trends to facilitate the adaptation of the health services to future needs and to identify research priorities
- The clarification of disease transmission to control the spread of contagious diseases, for example, by targeted vaccination programs.

Achievement of these aims is the prerequisite for the second perspective, which defines the scope of epidemiology from a public health point of view. Especially in this respect, the statement as given in Box 1.1 was issued by the IEA (International Epidemiological Association) conference already in 1975.

Box 1.1. Statement by the IEA conference in 1975 (White and Henderson 1976)

"The discipline of epidemiology, together with the applied fields of economics, management sciences, and the social sciences, provide the essential quantitative and analytical methods, principles of logical inquiry, and rules for evidence for:

- . . .;
- Diagnosing, measuring, and projecting the health needs of community and populations;
- Determining health goals, objectives and priorities;
- Allocating and managing health care resources;
- Assessing intervention strategies and evaluating the impact of health services."

This list may be complemented by the provision of tools for investigating consequences of disease as unemployment, social deprivation, disablement, and death.

1.2.2 Epidemiology in Relation to Other Disciplines

Biomedical, clinical, and other related disciplines sometimes claim that epidemiology belongs to their particular research area. It is therefore not surprising that biometricians think of epidemiology as a part of biometry and physicians define epidemiology as a medical science. On the one hand, biometricians have in mind that epidemiology uses statistical methods to investigate the distribution of health-related entities in populations as opposed to handling single cases. This perspective on distributions of events, conditions, etc. is statistics by its very nature. On the other hand, physicians view epidemiology primarily from a substantive view on diseases and their treatment. In doing so, each of them may disregard central elements that constitute epidemiology.

Moreover, as described at the beginning, epidemiology overlaps with various other domains that provide their methods and knowledge to answer epidemiological questions. For example, measurement scales and instruments to assess subjective well-being developed by psychologists can be applied by epidemiologists to investigate the psychological effects of medical treatments in addition to classical clinical outcome parameters. Social sciences provide indicators and methods of field work that are useful in describing social inequality in health, in investigating social determinants of health, and in designing population-based prevention strategies. Other examples are methods and approaches from demography that are used to provide health reports, from population genetics to identify hereditary factors, and from molecular biology to search for precursors of diseases and factors of susceptibility.

Of course, epidemiology does not only borrow methods from other sciences but has also its own methodological core. This pertains in particular to the development and adaptation of study designs. It is also true for statistical methods. In most cases they can directly be applied to epidemiological data, but sometimes peculiarities in the data structure may call for the derivation of special methods to cope with these requirements. This is in particular the case in genetic epidemiology when, for example, gene-environment interactions have to be modeled.

The line of demarcation between epidemiology and related disciplines is often blurred. Let us take clinical medicine as an example. In clinical practice, a physician decides case by case to diagnose and treat individual patients. To achieve the optimal treatment for a given subject, he/she will classify this patient and then make use of knowledge on the group to which the person belongs. This knowledge may come from randomized clinical trials but also from (clinical) epidemiological studies. A randomized clinical trial is a special case of a randomized controlled trial (RCT). In a broad sense, an RCT is an epidemiological experiment in which subjects in a population are randomly allocated to groups, that is, a study group where intervention takes place and a control group without intervention. This reveals the link between clinical and epidemiological studies, where the latter focus on populations while clinical trials address highly selected groups of patients. Thus, it may be controversial whether randomized clinical trials for drug approval (i.e., phase III trials) are to be considered part of epidemiology, but it is clear that a follow-up concerned with safety aspects of drug utilization (the so-called phase IV studies) needs pharmaco-epidemiological approaches.

When discussing the delimitation of epidemiology, the complex area of public health plays an essential role. According to Last's definition (Last 2001) public health deals with the health needs of the population as a whole, in particular the prevention and treatment of disease. More explicitly, "Public health is one of the efforts organized by society to protect, promote, and restore the people's health. It is the combination of sciences, skills, and beliefs that is directed to the maintenance and improvement of the health of all the people through collective or social actions. (…) Public health … goals remain the same: to reduce the amount of disease, premature death, and disease-produced discomfort and disability in the population. Public health is thus a social institution, a discipline, and a practice" (Last 2001). The practice of public health is based on scientific knowledge of factors influencing health and disease, where epidemiology is, according to Detels and Breslow (2002), "the core science of public health and preventive medicine" that is complemented by biostatistics and "knowledge and strategies derived from biological, physical, social, and demographic sciences."

In conclusion, epidemiology cannot be reduced to a subdivision of one of the contributing sciences, but it should be considered as a multidisciplinary science laying the grounds of the applied field of public health.

1.3 Historical Development of Epidemiology

In the following we highlight some milestones of the historical development of epidemiology. For a more comprehensive description of the history of epidemiology, we refer to chapter ▸History of Epidemiological Methods and Concepts of this handbook.

1.3.1 Historical Background

The word "epidemic," that is, something that falls upon people (ἐπί upon; δῆμος people), which was in use in ancient Greece, reflected already one of the basic ideas of modern epidemiology, namely, to look at diseases on the level of *populations*, or *herds* as they also have been called, especially in the epidemiology of infectious diseases. The link with the search for *causes* of illness was present in early writings of the Egyptians, Jews, Greeks, and Romans (Bulloch 1938). Both Hippocrates (ca 460 – ca 375 BC) and Galen (129 or 230–200 or 201) advanced etiological theories. The first stressed atmospheric conditions and “miasmata” but considered nutrition and lifestyle as well (Hippocrates 400 BC). The second distinguished three causes of an “epidemic constitution” in a population: an atmospheric one, susceptibility, and lifestyle. The basic book by Coxe (1846) contains a classification of Galen’s writings including the subject “etiology.” For a survey on the various editions of Galen’s work and a biography, see the essay by Siegel (1968).

Regarding more specific observations, the influence of dust in quarries on chronic lung diseases was mentioned in a Roman text of the first century. Paracelsus in 1534 published the first treatise on *occupational diseases*, entitled “Von der Bergsucht oder Bergkranckheiten drey Bücher, inn dreyzehen Tractat verfast unnd beschriben worden” (“On miners’ diseases”); see his biography in English by Pagel (1982). Ramazzini (1713) conjectured that the relatively high incidence of breast cancer among nuns was due to celibacy. Sixty-two years later, Percival Pott (1775) was among the first ones to phrase a comparative observation in quantitative terms. He reported that scrotal cancer was very frequent among London chimney sweeps and that their death rate due to this disease was more than 200 times higher than that of other workers.

The most prominent early observational epidemiological study is that of John Snow on cholera in London in 1853. He was able to record the mortality by this disease in various places of residence which he characterized by type of water supply and sewage system. By comparing morbidity and mortality rates, he concluded that polluted water was indeed the cause of cholera (Snow 1855).

Parallel to this emergence of observational epidemiology, three more currents of epidemiological thinking have emerged during the centuries and have interacted with each other as well as with the former, namely, (1) the debate on *contagion* and *living causal agents*, (2) *descriptive* epidemiology in the classical sense of health statistics, and (3) *clinical trials*.

A contagion can be suspected from recording cases and their location in time, space, families, and the like. The possibility of its involvement in epidemics has therefore no doubt been considered since time immemorial; it was alluded to the early writings mentioned at the beginning. Nevertheless, Hippocrates and Galen did not admit it. It played an important role in the thinking about *variolation* and later on *vaccination* as introduced by Jenner in 1796 (Jenner 1798). The essay by Daniel Bernoulli on the impact of variolation (Bernoulli 1766) was the beginning of the theory of *mathematical modeling* of the spread of diseases.

By contrast to a contagion itself, the existence of *living* pathogens cannot be deduced from purely epidemiological observations, but the discussion around it has often been intermingled with that about contagion and has contributed much to epidemiological thinking. Fracastoro (1521) wrote about a *contagium animatum*. In the sequel the idea came up again and again in various forms, for example, in the writings of Snow. It culminated in the identification of specific parasites, fungi, bacteria, and viruses as agents in the period from, roughly, 1840 when Henle, after Arabian predecessors dating back to the ninth century, definitely showed that mites cause scabies, until 1984 when HIV was detected.

As far as we know, the term "epidemiology" first appeared in Madrid in 1802. From the late nineteenth century to about the middle of the twentieth, it was restricted to *epidemical infectious* diseases until it took its present meaning (see Sect. 1.2.1 and Greenwood 1932).

Descriptive epidemiology had various precursors, mainly in the form of church and military records on the one hand (Marshall and Tulloch 1838), and life tables on the other (Graunt 1662; Halley 1693). In the late eighteenth century, local medical statistics started to appear in many European cities and regions. They took a more systematic turn with the work of William Farr (1975). This lasted from 1837 when he was appointed to the General Register Office in London until his retirement in 1879. In particular, he developed classifications of diseases that led to the first International List of Causes of Death, to be adopted in 1893 by the International Statistical Institute. Farr took also part in the activities of the London Epidemiological Society, founded in 1850 with him and Snow as founding members and apparently the oldest learned society featuring the word "epidemiological" in its name.

Geographical epidemiology, that is, the presentation of health statistics in the form of maps, also started in the nineteenth century (Rupke 2000).

If we mean by a clinical trial a *planned*, *comparative*, and *quantitative* experiment on humans in order to learn something about the efficacy of a curative or preventive treatment in a clinical setting, James Lind is considered having done the first one. In 1747 he tried out six different supplements to the basic diet of 12 sailors suffering from scurvy and found that citrus fruits, and only these, cured the patients (Lind 1753). Later he also compared quinine to treat malaria with less well-defined control therapies (Lind 1771).

The first more or less rigorous trial of a preventive measure was performed by Jenner with 23 vaccinated people. For his trial he used what is now being called "historical controls," that is, he compared these vaccinated people with unvaccinated ones of the past who had not been specially selected beforehand for the purpose of the trial (Jenner 1798).

In the nineteenth century, some physicians began to think about the general principles of clinical trials and already emphasized probabilistic and statistical methods (Louis 1835; Bernard 1865). Some trials were done, for example, on the efficacy of bloodletting to treat pneumonia, but rigorous methods in the modern sense were established only after World War II, beginning in 1948 with the pioneer trial on the treatment of pulmonary tuberculosis by streptomycin as described in Hill (1962).

Let us conclude this all too short historical sketch with a few remarks on the history of applications of epidemiology:

Clinical trials have always been tied, by their very nature, to immediate applications as in the above-mentioned examples; hence we will not dwell on this further.

Observational epidemiology, including classical descriptive epidemiology, has led to hygienic measures. In fact, coming back to a concept of Galen (1951), one might define *hygiene* in a modern and general sense as applied observational epidemiology, its task being to diminish or to eliminate causal factors of any kind.

For example, the results of Snow's study on cholera found rapid applications in London but not in places like Hamburg where 8,600 people died in the cholera epidemic of 1892.

Hygiene was a matter of much debate and activity during the entire nineteenth century, although, before the identification of living pathogens, most measures taken were necessarily not directed against a known specific agent, with the exception of *meat inspection* for trichinae. This was made compulsory in Prussia in 1875 as proposed by Rudolf Virchow, one of the pioneers of modern hygiene and also an active politician (Ackerknecht 1953).

Hygienic activities generally had their epidemiological roots in the descriptive health statistics mentioned above. These statistics usually involved only factors like time, place of residence, sex, and age, but Virchow, for example, analyzed during the years 1854–1871 the mortality statistics for the city of Berlin and related mortality to social factors like poverty, crowded dwellings, and dangerous professions, thus becoming a forerunner of *social epidemiology*.

As a result of such reflections as well as of political pressure, large *sewage systems* were built in Europe and North America, the *refuse disposal* was re-organized, and the *water supply* improved. Other hygienic measures concerned the structure and functioning of *hospitals*, from reducing the number of patients per room and dispersing wards in the form of pavilions to antiseptic rules. The latter had mainly been inspired by more or less precise epidemiological observations on infections after the treatment of wounds and amputations (Tenon 1788; Simpson 1868–1869, 1869–1870; Ackerknecht 1967) and on puerperal fever (Gordon 1795; Holmes 1842–1843; Semmelweis 1861). *Nosocomial infections* are still a major public health problem and fighting them a major application of epidemiology.

1.3.2 Milestones in Epidemiological Research

The initiation of numerous epidemiological studies after the Second World War accelerated the research in this field and led to a systematic development of study designs and methods. In the following some exemplary studies are introduced that served as role models for the design and analysis of many subsequent investigations. It is not our intention to provide an exhaustive list of all major studies since that time, if at all feasible, but to exhibit some cornerstones marking the most important steps in the evolution of this science. Each of them had its own peculiarities with a high impact both on methods and epidemiological reasoning as well as on health policies.

The usefulness of descriptive study designs has been convincingly demonstrated by migrant studies comparing the incidence or mortality of a disease within a certain population between the country of origin and the new host country. Such observations offer an exceptional opportunity to distinguish between potential contributions of genetics and environment to the development of disease and thus make it possible to distinguish between the effects of nature and nurture. The most prominent examples are provided by investigations on Japanese migrants to Hawaii and California. For instance, the mortality from stomach cancer was much higher in Japan than among US inhabitants, whereas for colon cancer the relationship was reverse. Japanese migrants living in California had a mortality pattern that lay between these two populations. It was thus concluded that dietary and other lifestyle factors had a stronger impact than hereditary factors, which was further supported by the fact that the sons of Japanese immigrants in California had an even lower risk for stomach cancer and a still higher risk for cancer of the colon than their fathers (Buell and Dunn 1965).

One of the milestones in epidemiological research was the development of rigorous case-control designs, which facilitate the investigation of risk factors for chronic diseases with long induction periods. The most famous study of this type, although not the first one, is the study on smoking and lung cancer by Doll and Hill (1950). As early as in 1943, the German pathologist Schairer from the Scientific Institute for Research into the Hazards of Tobacco, Jena, published together with Schöniger a case-control study comparing 109 men and women deceased from lung cancer with 270 healthy male controls as well as with 318 men and women who died from other cancers with regard to their smoking habits (Schairer and Schöniger 1943). Judged by modern epidemiological standards, this study had several weaknesses; still, it showed a clear association of tobacco smoking and lung cancer. The case-control study by Doll and Hill was much more sophisticated in methodological terms. Over the whole period of investigation from 1948 to 1952, they recruited 1,357 male and 108 female patients with lung cancer from several hospitals in London and matched them with respect to age and sex to the same number of patients hospitalized for non-malignant conditions. For each patient, a smoking history was collected. Without going into detail here, these data supported a strong positive association between smoking and lung cancer. Despite the methodological concerns regarding case-control studies, Doll and Hill believed that smoking was responsible for the development of lung cancer. The study became a landmark that inspired future generations of epidemiologists to use this methodology (cf. chapter ▸Case-Control Studies of this handbook). It remains a model for the design and conduct of case-control studies until today, with excellent suggestions on how to reduce or eliminate selection, interview, and recall bias.

Because of the strong association observed in their case-control study, they initiated a cohort study of 20,000 male British physicians in 1951, known as the British Doctors' Study. These were followed over time to investigate the association between smoking and lung cancer prospectively. The authors compared mortality from lung cancer among those who never smoked with that among all smokers and with those who smoked various numbers of cigarettes per day (Doll and Hill 1954, 1964; Doll and Peto 1978).

Another, probably even more important cohort study was the Framingham Heart Study that was based on the population of Framingham, a small community in Massachusetts. The study was initiated in 1949 to yield insights into causes of cardiovascular diseases (CVD) (see chapter ▸Cardiovascular Health and Disease of this handbook). For this purpose, 5,127 participants free from coronary heart disease (CHD), 30 to 59 years of age, were examined and then observed for nearly 50 years to determine the rate of occurrence of new cases among persons free of disease at first observation (Dawber et al. 1951; Dawber 1980). The intensive biennial examination schedule, long-term continuity of follow-up and investigator involvement, and incorporation of new design components over its decade-long history have made this a uniquely rich source of data on risk factors for CVD. The study served as a reference and good example for many subsequent cohort studies in this field which adopted its methodology. In particular, analysis of the Framingham data led to the development of the perhaps most important modeling technique in epidemiology, the multiple logistic regression (Truett et al. 1967; see chapter ▶Regression Methods for Epidemiological Analysis of this handbook).

Two other leading examples of cohort studies conducted within a single population or for comparison of multiple populations to assess risk factors for cardiovascular events are the Whitehall Study of British civil servants (Rose and Shipley 1986; see also chapter ▶Occupational Epidemiology of this handbook) and the Seven Countries Study of factors accounting for differences in CHD rates between populations of Europe, Japan, and North America (Keys 1980; Kromhout et al. 1995; see chapter ▶Cardiovascular Health and Disease of this handbook).

In contrast to the above cohort studies that focused on cardiovascular diseases, the US Nurses' Health Study is an impressive example of a multipurpose cohort study. It recruited over 120,000 married female nurses, 30 to 55 years of age, in a mail survey in 1976. In this study, information on demographic, reproductive, medical, and lifestyle factors was obtained. Nurses were contacted every 2 years to assess outcomes that occurred during that interval and to update and to supplement the exposure information collected at baseline. Various exposure factors like the use of oral contraceptives, postmenopausal hormone therapy, and fat consumption were related to different outcomes such as cancer and cardiovascular disease (Lipnick et al. 1986; Willett et al. 1987; Stampfer et al. 1985). Recent results had an essential impact on the risk-benefit assessment of postmenopausal hormone therapy speaking against its use over extended periods (Chen et al. 2002).

Final the proof of a causal relationship is provided by experimental studies, namely, intervention trials. The most famous and largest intervention trial was the so-called Salk vaccine field trial in 1954 where nearly one million school children were randomly assigned to one of the two groups, a vaccination group that received the active vaccine and a comparison group receiving placebo. A 50% reduction of the incidence of paralytic poliomyelitis was observed in the vaccination group as compared to the placebo group. This gave the basis for the large-scale worldwide implementation of poliomyelitis vaccination programs for disease prevention.

In recent years, the accelerated developments in molecular biology were taken up by epidemiologists to measure markers of exposure, early biological effects, and host characteristics that influence response (susceptibility) in human cells, blood, tissue, and other material. These techniques augment the standard tools of epidemiology in the investigation of low-level risks, risks imposed by complex exposures, and the modification of risks by genetic factors. The use of such biomarkers of exposure and effect has led to a boom of the so-called molecular epidemiology (Schulte and Perera 1993; Rothman et al. 2012; Toniolo et al. 1997; chapter ▸Molecular Epidemiology of this handbook), a methodological approach with early origins. These developments were accompanied by the sequencing of the human genome and the advances in high-throughput genetic technologies that led to the rapid progress of genetic epidemiology (Khoury et al. 1993; chapter ▸Statistical Methods in Genetic Epidemiology of this handbook). The better understanding of genetic factors and their interaction with each other and with environmental factors in disease causation is a major challenge for future research. Here, large-scale studies, often accomplished by the pooling of several studies, paved the way for genome-wide association studies (GWAS) assessing genetic variants that entail increased disease risks.

1.3.3 Methodological Limits

The successes of epidemiology in identifying major risk factors of chronic diseases have been contrasted with more subtle risks that epidemiologists have seemingly discovered like the association between environmental tobacco smoke exposure and lung cancer. Such risks are difficult to determine and false alarms may result from chance findings. Thus it is not surprising that many studies showed conflicting evidence, that is, some studies seem to reveal a significant association with a suspected risk factor while others do not. The unreflected publication of such contradictory results in the lay press leads to opposing advice and thus to an increasing anxiety in the public. This has given rise to a controversial debate about the methodological weaknesses of epidemiology that culminated in the article “Epidemiology faces Its Limits” by Taubes (1995) and the discussions that it prompted.

In investigating low relative risks, say, below 2 or even below 1.5, the methodological shortcomings inherent in observational designs become more serious. Such studies are more prone to yield false positive or false negative findings due to the distorting effects of misclassification, bias, and confounding (see chapters ▸Misclassification, ▸Confounding and Interaction, ▸Sensitivity Analysis and Bias Analysis, and ▸Measurement Error of this handbook). For instance, the potential effect of environmental tobacco smoke (ETS) on lung cancer was denied because misclassification of only a few active smokers as non-smokers would result in relative risks that might explain all or most of the observed association between ETS and the risk of lung cancer in non-smokers (Lee and Forey 1996). Validation studies showed that this explanation was unlikely (Riboli et al. 1990; Wells et al. 1998). Thus, the numerous positive findings and the obvious biological plausibility of the exposure-disease relationship support the conclusion of a harmful effect of ETS (Boffetta et al. 1998; Chan-Yeung and Dimich-Ward 2003; IARC Monograph on ETS 2004). This example also illustrates that the investigation of low relative risks is not an academic exercise but may be of high public health relevance if a large segment of the population is exposed.

It is often believed that large-scale studies are needed to identify small risks since such studies result in narrower confidence intervals. However, a narrow confidence interval does not necessarily mean that the overwhelming effects of misclassification, bias, and confounding are adequately controlled by simply increasing the size of a study. Even sophisticated statistical analyses will never overcome serious deficiencies of the database. The fundamental quality of the data collected or provided for epidemiological purposes is therefore the cornerstone of any study and needs to be prioritized throughout its planning and conduct (see chapter ▶Quality Control and Good Epidemiological Practice of this handbook). In addition, the refinement of methods and measures involving all steps from study design over exposure and outcome assessment to the final data analysis, incorporating, for example, molecular markers, may help to push the edge of what can be achieved with epidemiology a little bit further.

Nevertheless a persistent problem is "The pressures to publish inconclusive results and the eagerness of the press to publicize them ..." (Taubes 1995). This pressure to publish positive findings that are of questionable validity imposes a particular demand on researchers not only to report and interpret study results carefully in peer-reviewed journals but also to communicate potential risks to the lay press in a comprehensible manner that accounts for potential limitations. Both authors and editors have to take care that the pressure to publish does not lead to a publication bias favoring positive findings and dismissing negative results.

1.4 Concepts, Methodological Approaches, and Applications

1.4.1 Concepts

Epidemiology may be considered as minor to physical sciences because it does not investigate the biological mechanism leading from exposure to disease as, for example, toxicology does. However, the ability of identifying modifiable conditions that contribute to health outcomes without identifying the biological mechanisms or the agent(s) that lead to these outcomes is a strength of epidemiology: It is not always necessary to wait until the underlying mechanism is completely understood in order to facilitate preventive action. This is illustrated by the historical examples of the improvements of environmental hygiene that led to a reduction of infectious diseases like cholera that was possible before the identification of *vibrio cholerae*.

What distinguishes epidemiology is its perspective on groups or populations rather than individuals. It is this demographic focus where statistical methods enter the field and provide the tools needed to compare different characteristics relating to disease occurrence between populations. Epidemiology is a comparative discipline that contrasts occurrence of diseases and characteristics between different time periods, different places, or different groups of persons. The comparison of groups is a central feature of epidemiology, be it the comparison of morbidity or mortality in populations with and without a certain exposure or the comparison of exposure between diseased subjects and a control group. Inclusion of an appropriate reference group (non-exposed or non-diseased) for comparison with the group of interest is a condition for causal inference.

In experimental studies efficient means are available to minimize the potential for bias. Due to the observational nature of the vast majority of epidemiological studies, bias and confounding are the major problems that may restrict the interpretation of the findings if not adequately taken into account (see chapters ▸Confounding and Interaction, ▸Sensitivity Analysis and Bias Analysis, and ▸Design and Planning of Epidemiological Studies of this handbook). Although possible associations are analyzed and reported on a group level, it is important to note that only data that provide the necessary information on an individual level allow the adequate consideration of confounding factors (see chapter ▸Descriptive Studies of this handbook).

Most epidemiological studies deal with mixed populations. On the one hand, the corresponding heterogeneity of covariables may threaten the internal validity of a study, because the inability to randomize study subjects in observational studies may impair the comparability between study groups due to confounding. On the other hand the observation of natural experiments in diverse groups of individuals enables epidemiologists to make statements about the real world and thus contributes to the external validity of the results. This population perspective enables the assessment of effectiveness of a treatment rather than its efficacy when evaluating interventions.

Due to practical limitations, it may not be feasible to obtain a representative sample of the whole population of interest in a given study. It may even be desired to investigate only defined subgroups of a population. Whatever the reason, a restriction on subgroups may not necessarily impair the meaning of the obtained results, but it may increase the internal validity of a study. Thus, it is a misconception that the cases always need to be representative of all persons with the disease and that the reference group always should be representative of the general non-diseased population. What is important is a precise definition of the population base, that is, in a case-control study, cases and controls need to originate from the same source population and the same inclusion/exclusion criteria need to be applied to both groups. This implies that the interpretation of study results has to consider possible limitations with respect to their generalizability beyond the source population.

Rarely a single positive study will provide sufficient evidence to justify an intervention. Limitations inherent in most observational studies require the consideration of alternative explanations of the findings and confirmation by independent evidence from other studies in different populations before preventive action can be recommended with sufficient certainty. The interpretation of negative studies deserves the same scrutiny as the interpretation of positive studies. Negative results should not hastily be interpreted to prove an absence of the association under investigation (Doll and Wald 1994). Besides chance, false-negative results may easily be due to a weak design and conduct of a given study, that is, absence of evidence is not evidence of absence (Altman and Bland 1995).

1.4.2 Study Designs

Epidemiological reasoning consists of three major steps. First, a statistical association between an explanatory characteristic (exposure) and the outcome of interest (disease) is assessed. Then, depending on the pattern of the association,

Table 1.1 Types of epidemiological studies

Type of study	Alternative name	Unit of study
Observational		
Ecological study	Correlational study	Populations
Cross-sectional study	Prevalence study; survey	Individuals
Case-control study	Case-referent study	Individuals
Cohort study	Follow-up study	Individuals
Experimental	*Intervention studies*	
Community trial	Community intervention study	Communities
Field trial	–	Healthy individuals
Randomized controlled trial	RCT	Individuals
Clinical trial	Therapeutic study[a]	Individual patients

[a] Clinical trials are included here since conceptually they are linked to epidemiology, although they are often not considered as epidemiological studies. Clinical trials have developed into a vast field of its own because of methodological reasons and their economic importance.

a hypothetical (biological) inference about the disease mechanism is formulated that can be refuted or confirmed by subsequent studies. Finally, when a plausible conjecture about the causal factor(s) leading to the outcome has been acknowledged, alteration or reduction of the putative cause and subsequent observation of the resulting disease frequency provide the verification or refutation of the presumed association.

In practice these three major steps are interwoven in an iterative process of hypothesis generation by descriptive and exploratory studies, statistical confirmation of the presumed association by analytical studies, and, if feasible, implementation and evaluation of intervention activities, that is, experimental studies. An overview of the different types of study and some common alternative names are given in Table 1.1.

A first observation of a presumed relation between exposure and disease is often done at the group level by correlating one group characteristic with an outcome, that is, in an attempt to relate differences in morbidity or mortality of population groups to differences in their local environment, culture, lifestyle, or other factors. Such correlational studies that are usually based on existing data (see chapter ▶Use of Health Registers of this handbook) are prone to the so-called ecological fallacy since the compared populations may differ in factors other than the ones of interest that are related to the disease. For a detailed discussion of this issue, we refer to the chapter ▶Descriptive Studies of this handbook. Nevertheless, ecological studies can provide clues to etiological factors and may serve as a gateway towards more detailed investigations. In subsequent analytical studies the investigator determines whether the relationship in question is also present among individuals, either by asking whether persons with the disease have the characteristic more frequently than those without the disease or by asking whether persons with the characteristic develop the disease more frequently than those not having it. The investigation of an association at the individual level is considered to be less vulnerable to be mixed up with the effect of a third common factor.

Studies that are primarily designed to describe the distribution of existing variables that can be used for the generation of broad hypotheses are often classified as descriptive studies (cf. chapter ▸Descriptive Studies of this handbook). Analytical studies examine an association, that is, the relationship between a risk factor and a disease in detail, and conduct a statistical test of the corresponding hypothesis. Typically the two main types of epidemiological studies, that is, case-control study and cohort study, belong to this category (see chapters ▸Cohort Studies and ▸Case-Control Studies of this handbook). However, a clear-cut distinction between analytical and descriptive study designs is not possible. A case-control study may be designed to explore associations of multiple exposures with a disease. Such “fishing expeditions” may better be characterized as descriptive rather than analytical studies. A cross-sectional study is descriptive when it surveys a community to determine the health status of its members. It is analytical when the association of an acute health event with a recent exposure is analyzed.

Cross-sectional studies provide descriptive data on prevalence of diseases useful for health-care planning. Prevalence data on risk factors from descriptive studies also help in planning an analytical study, for example, for sample size calculations. The design is particularly useful for investigating acute effects but has significant drawbacks in comparison to a longitudinal design because the temporal sequence between exposure and disease usually cannot be assessed with certainty, except for invariant characteristics like genetic polymorphisms. In addition, it cannot identify incident cases of a chronic disease (see chapter ▸Descriptive Studies of this handbook).

Both case-control and cohort studies are in some sense longitudinal because they incorporate the temporal dimension by relating exposure information to time periods that are prior to disease occurrence. These two study types – in particular when data are collected prospectively – are therefore more informative with respect to causal hypotheses than cross-sectional studies because they are less prone to the danger of “reverse causality” that may emerge when information on exposure and outcome relates to the same point in time. The best means to avoid this danger are prospective designs where the exposure data are collected prior to disease. Typically these are cohort studies, either concurrent or historical, as opposed to retrospective studies, that is, case-control studies where information on previous exposure is collected from diseased and non-diseased subjects. For further details of the strengths and weaknesses of the main observational designs, see chapter ▸Design and Planning of Epidemiological Studies of this handbook.

The different types of studies are arranged in Table 1.1 in ascending order according to their ability to support the causality of a supposed association. The criteria summarized by Hill (1965) have gained wide acceptance among epidemiologists as a checklist to assess the strength of the evidence for a causal relationship. However, an uncritical accumulation of items from such a list cannot replace the critical appraisal of the quality, strengths, and weaknesses of each study. The weight of evidence for a causal association depends in the first place on the type of study, with intervention studies providing the strongest evidence (Table 1.2)

Table 1.2 Reasoning in different types of epidemiological study

Study type	Reasoning
Ecological	Descriptive; association on group level may be used for development of broad hypotheses
Cross-sectional	Descriptive; individual association may be used for development and specification of hypotheses
Case-control	Increased prevalence of risk factor among diseased as compared to non-diseased may indicate a causal relationship
Cohort	Increased risk of disease among exposed as compared to non-exposed indicates a causal relationship
Intervention	Modification (reduction) of the incidence rate of the disease confirms a causal relationship

(see chapter ▶Intervention Trials of this handbook). The assessment of causality has then to be based on a critical judgment of evidence by conjecture and refutation (see chapter ▶Basic Concepts of this handbook for a discussion of this issue).

1.4.3 Collection of Data

Data are the foundation of any empirical study. It is fundamental to avoid any sort of systematic bias in the planning and conduct of an epidemiological study, be it information or selection bias. Errors that have been introduced during data collection can in most cases not be corrected later on. Exceptions from this rule are, for example, measurement instruments yielding distorted measurements where size and direction of the systematic error is known so that the individual measurement values can be calibrated accordingly. In other instances statistical methods are offered to cope with measurement error (see chapter ▶Measurement Error of this handbook). However, such later efforts are second choice and an optimal quality of the original data must be the primary goal. Selection bias may even be worse as it cannot be controlled for and may affect both the internal and the external validity of a study. Standardized procedures to ensure the quality of the original data to be collected for a given study are therefore crucial (see chapter ▶Quality Control and Good Epidemiological Practice of this handbook).

Original data will usually be collected by questionnaires, the main measurement instrument in epidemiology. Epidemiologists have neglected for a long time the potential in improving the methods for interviewing subjects in a highly standardized way and thus improving the validity and reliability of this central measurement tool. Only in the late 1990s, it has been recognized that major improvements in this area are not only necessary but also possible, for example, by adopting methodological developments from the social sciences (Olsen et al. 1998). Chapter ▶Epidemiological Field Work in Population-Based Studies of this handbook is devoted to the basic principles and approaches in this field. Prior to the increased awareness related to data collection methods, the area of exposure assessment has developed into a flourishing research field that provided advanced tools and guidance for researchers (White et al. 2008; Kromhout 1994; Ahrens 1999; Nieuwenhuijsen 2003; chapter ▶Exposure Assessment of this handbook). Recent advances in molecular epidemiology have introduced new possibilities for exposure measurement that are now being used in addition to the classical questionnaires. However, since the suitability of biological markers for the retrospective assessment of exposure is limited due to the short half-life of most agents that can be examined in biological specimens, the use of interviews will retain its importance but will change its face. Computer-aided data collection with built-in plausibility checks – more and more being conducted in the form of telephone interviews or even by the Internet – will partially replace the traditional paper and pencil approach (see chapter ▶Internet-Based Epidemiology of this handbook).

Often it may not be feasible to collect primary data for the study purpose due to limited resources or because of the specific research question. In such cases, the epidemiologist can sometimes exploit existing databases such as registries (see chapter ▶Use of Health Registers of this handbook). Here, he/she usually has to face the problem that such "secondary" or "routine" data may have been collected for administrative or other purposes. Since such data are collected on a routine basis without the claim for subsequent systematic analyses, they may be of limited quality. Looking at these data from a research perspective thus often reveals inconsistencies that had not been noticed before. The degree of standardization that can be achieved in collection, documentation, and storage is particularly low if personnel of varying skills and levels of training is involved. Moreover, changes in procedures over time may introduce additional systematic variation. Measures for assessing the suitability and quality of the data and for careful data cleaning are therefore of special importance.

The scrutiny, time, and effort that need to be devoted to any data use - be it routine data or newly collected data - before they can be used for analysis are rarely addressed in standard textbooks of epidemiology and often neglected in study plans. This is also true for the coding of variables like diseases, pharmaceuticals, or job titles. They deserve special care with regard to training and quality assurance. Regardless of all efforts to ensure an optimal quality during data collection, a substantial input is needed to guarantee standardized and well-documented coding, processing, and storing of data (see chapter ▶Data Management in Epidemiology of this handbook). Residual errors that remain after all preceding steps need to be scrutinized and, if possible, corrected (see chapter ▶Quality Control and Good Epidemiological Practice of this handbook). Sufficient time has to be allocated for this work package that precedes the statistical analysis and publication of the study results. Finally, all data and study materials have to be stored and documented in a fashion that allows future use and/or sharing of the data or auditing of the study. Materials to be archived should not only include the electronic files of raw data and files for the analyses but also the study protocol, computer programs, the documentation of data processing and data correction, measurement protocols, and the final report. Both, during the conduct of the study as well as after its completion, materials and data have to be stored in a physically safe place with limited access to ensure safety and confidentiality even if the data have been anonymized.

1.4.4 Application of Epidemiological Knowledge

Epidemiological knowledge concerns *populations*. There are two ways to use this knowledge. The first is *group-oriented*: It consists in applying knowledge about a specific population directly to this population itself. This is part of *public health*. The conceptually simplest applications of this kind concern the planning of the health system (chapter ▸Health Services Research of this handbook) and of health strategies. For instance, epidemiological studies have shown that people exposed to inhaling asbestos fibers are prone to develop asbestosis and its sequels like cor pulmonale. We apply this knowledge to the entire population by prohibiting the use of asbestos.

The second path is taken when we are confronted with an individual person, typically in a clinical setting: We can then regard this person as a member of a population for which relevant epidemiological knowledge is available and deal with her or him accordingly. As an example, a physician confronted with a child suffering from medium dehydration due to acute diarrhea, knows from clinical trials that oral rehydration will normally be a very efficient treatment. Hence she/he will apply it in this particular case.

Clinical epidemiology plays a major role for the second path, where epidemiological knowledge is applied in all phases of clinical decision making, that is, in daily clinical practice, starting with diagnosis, passing to therapy, and culminating in prognosis and advice to the patient – including individual preventive measures.

1.4.4.1 Prevention

The first of the two preceding examples belongs to *population-based prevention* (see chapters ▸Intervention Trials, ▸Community-Based Health Promotion, and ▸Epidemiology in Developing Countries of this handbook). The underlying idea is to diminish the influence of risk factors identified by previous observational epidemiological studies. Risks of transmission of infectious diseases have long played a particular role in public health: Their influence was reduced by public hygiene in the classical sense. Applying observational epidemiology in order to diminish or eliminate risk factors has therefore been termed *hygiene* in a modern, general sense. Preventive measures in this context are sometimes themselves subject to an a posteriori evaluation which may bear on the one hand on the way they have been implemented and on the other hand on their effectiveness.

Population-based preventive measures can also be derived from results of experimental epidemiology. The most important applications of this kind are *vaccinations* performed systematically within a given population. They have to be subjected to rigorous efficacy trials before implementation. *Preventive drug treatments*, for example, against malaria or cardiovascular events, fall into the same category.

In many cases the target population itself is determined by a previous epidemiological study. For instance, dietary recommendations to reduce cardiovascular problems, and vaccinations against Hepatitis B, yellow fever, or influenza, are primarily given to people that were identified as being at *high risk* to contract the disease in question.

1.4.4.2 Screening

Population-based *treatments* as a measure of public health are conceivable but hardly ever implemented. There exists, however, a population-based application of epidemiology in the realm of *diagnosis*, namely, screening (see chapter ▸Screening of this handbook). Its purpose is to find yet unrecognized cases of disease or health defects by appropriate tests that can be rapidly executed within large population groups. The ultimate aim is mostly to allow a treatment at an early stage. Occasionally, screening was also performed in order to isolate infected people, for example, lepers. Classical examples of screening include mass X-ray examination to detect cases of pulmonary tuberculosis or breast cancer and cytological tests to identify cancer of the cervix uteri. Screening programs may concentrate on *high-risk groups* if it would be unfeasible, too expensive, or too dangerous to examine the entire original population. A striking example is the screening for pulmonary tuberculosis in Norway where most of the prevalent cases were found at an early stage by systematic X-ray examinations of only a small fraction of the population (Harstad et al. 2009).

1.4.4.3 Case Management

The concept of the *individual* risk of a person (see chapters ▸Rates, Risks, Measures of Association and Impact and ▸Cohort Studies of this handbook) that underlies the definition of risk groups represents a particular case of the second way of applications of epidemiology, that is, dealing with an individual person on the basis of epidemiological knowledge about populations which he/she is deemed to belong to. The most important application of this idea, however, is *clinical epidemiology* which was also called *statistics in clinical medicine*. It is the art of case management in the most general sense: diagnosis using the epidemiological characteristics of medical tests like sensitivity and specificity, treatment using the results of clinical trials, and prognosis for a specific case based again on relevant epidemiological studies. Chapter ▸Clinical Epidemiology and Evidence-Based Health Care of this handbook describes in detail fairly sophisticated procedures involving all aspects of case management including the opinion of the patient or his/her relatives and considerations of cost, secondary effects, and quality of life as elements entering the therapeutic decision.

1.4.4.4 Health Services

Health services research (HSR) is a vast and multiform field. It has no concise and generally accepted definition, but still there is a more or less general agreement about its essential ideas. Its purpose is to lay the general, scientific foundations for health policy in order to improve the health of people as much as possible under the constraints of society and nature (cf. chapter ▸Health Services Research of this handbook). The subjects of HSR are, in the first place, the underlying *structures*, that is, the basic elements concerned by questions of health and the relations between them; in the second place the *processes* of health-care delivery; and in the third place the *effects* of health services on the health of the general population.

On the methodological side, HSR means *analysis* and *evaluation* of all of these aspects. The tools are mainly coming from mathematics and statistics, economics, and sociology together with knowledge from clinical medicine and basic sciences like biology. Epidemiology plays a particularly important role. Evaluation implies *comparison*: comparison of different existing health-care systems and comparison of an existing one with theoretical, ideal systems in order to design a better one. The comparison of health-care systems of different countries has been a favorite subject. One of the main "factors" that distinguishes them is the way medical services are being paid for and the form of health insurance.

The basic elements are physicians, nurses and other personnel, hospitals, equipment, and money, but also the population served by the health system. Relations between these elements comprise *health needs* and *access to services* but also the *organization* of the health system.

Processes of health-care delivery may of course mean the usual clinical curative treatment of patients but also person- or community-based preventive actions including environmental measures, health education, or health strategies like the one that led to the eradication of smallpox.

Finally, the effects of health services, that is, the *output*, can be measured in many ways, for example, by morbidity and mortality, life expectancy, quality of life preserved or restored, and economic losses due to illness. However, the quantitative assessment of *effectiveness*, that is, the value of outputs relative to (usually monetary) inputs, is still in the limelight.

Epidemiology as a method serves two purposes. On the one hand, the results of epidemiological investigations enter the field as basic parameters. Some experimental epidemiological studies like intervention trials are even considered as *belonging* to health services research. On the other hand, many methods used in health services research that stem from mathematical statistics and whose goal is to study the influence of various factors on outcomes are formally the same as those employed in epidemiology. Given the importance and complexity of the subject, many different "approaches" and "models" have been proposed and tried out. Earlier ones were still fairly descriptive and static, focusing on the functioning of the health services or on health policy with a strong emphasis on the economic aspects. The input-output model where the effects of changes of essential inputs on the various outcomes of interest are studied, if possible in a quantitative way, is more recent. More than others it allows to a large extent a "modular" approach, separating from each other the investigation of different parts or levels of the health services.

1.4.5 Ethical Aspects

The protection of human rights is one of the most crucial aspects of all studies on humans. Although there are substantial differences between experimental and observational studies, they both have to face the challenging task to protect the privacy of all individuals taking part in a study. This also implies as a basic principle that study subjects are asked for their informed consent.

Another ethical angle of epidemiological research concerns the study quality. Poorly conducted research may lead to unsubstantiated and wrong decisions in clinical practice or policy making in public health and may thus cause harm to individuals but also to society as a whole. Therefore, guidelines such as the "Good Epidemiological Practice" provided by the International Epidemiological Association in 1998 have been prepared to maintain high study quality and to preserve human rights.

Of course, the four general principles of the Declaration of Helsinki (World Medical Association 2000) have to be followed, that is, autonomy (respect for individuals), beneficence (do good), non-maleficence (do no harm), and justice. These principles are of particular relevance in randomized controlled trials, where the intervention (or non-intervention) may involve negative consequences for participants.

Various recent developments in epidemiological research constitute a new challenge regarding ethical requirements. First, automated record linkage databases are now at least partly available that capture both exposure and outcome data on an individual level. Such databases have raised questions about confidentiality of patient's medical records, authorizing access to person-specific information, and their potential misuse. Second, the inclusion of molecular markers in epidemiological studies has led to a controversial debate on the potential benefit or harm of results gained by genetic and molecular epidemiological studies. This raises the following questions: Can knowledge on genetic markers be used in primary prevention programs? How should this knowledge be communicated to the study subjects who may be forced into the conflict between their individual "right to know" and their "right not to know?" A third driver of ethical questioning has been the discussion about integrity and conflict of interests, in particular when epidemiological studies are sponsored by interest groups or when the results are contradictory.

As a consequence, an increasing awareness that ethical conduct is essential to epidemiological research can be observed among epidemiologists. Thus, it is not surprising that now basic principles of integrity, honesty, truthfulness, fairness and equity, respect for people's autonomy, distributive justice, and doing good and not harm have become an integral part in the planning and conduct of epidemiological studies. Chapter ▶Ethical Aspects of Epidemiological Research of this handbook is devoted to all of these aspects.

1.5 Statistical Methods in Epidemiology

The statistical analysis of an empirical study relates to all its phases. It starts with the planning phase where ideally all details of the subsequent analysis should be laid down (see chapter ▶Design and Planning of Epidemiological Studies of this handbook). This concerns defining the variables to be collected and their scale; the methods how they should be summarized, for example, via means, rates, or odds; the appropriate statistical models to be used to capture the relationship between exposures and outcomes; the formulation of the research questions as statistical hypotheses; the calculation of the necessary sample size based on a given power or vice versa the power of the study based on a fixed sample size; and the appropriate techniques to check for robustness and sensitivity. It is crucial to have in mind that the study should be planned and carried out in such a way that its statistical analysis is able to answer the research questions we are interested in. If the analysis is not already adequately accounted for in the planning phase or if only a secondary analysis of already existing data can be done, the results will probably be of limited validity and interpretability.

1.5.1 Principles of Data Analysis

Having collected the data, the first step of a statistical analysis is devoted to the cleaning of the dataset. Questions to be answered are as follows: "Are the data free of measurement or coding errors?" "Are there differences between centers?" "Are the data biased, already edited, or modified in any way?" "Have data points been removed from the dataset?" "Are there outliers or internal inconsistencies in the dataset?" A sound and thorough descriptive analysis enables the investigator to inspect the data. Cross-checks based, for example, on the range of plausible values of the variables and cross tabulations of two or more variables have to be carried out to find internal inconsistencies and implausible data (see chapter ▶Data Management in Epidemiology of this handbook). Graphical representations such as scatterplots, boxplots, and stem-and-leaf diagrams help to detect outliers and irregularities. Calculating various summary statistics such as *mean* as compared to *median*, *standard deviation* as compared to *median absolute deviation from the median* is also reasonable to reveal deficiencies in the data. Special care has to be taken to deal with measurement errors and missing values. In both cases, special statistical techniques are available to cope with such data (see chapters ▶Measurement Error and ▶Missing Data of this handbook).

After having cleaned the dataset, descriptive measures such as cross tabulations, correlation coefficients or graphical representations like boxplots will help the epidemiologist to understand the structure of the data. Such summary statistics need, however, to be interpreted carefully. They are descriptive by their very nature and are not to be used to formulate statistical hypotheses that are subsequently investigated by a statistical significance test based on the same dataset.

In the next step parameters of interest such as relative risks or incidence rates should be estimated. The calculated point estimates should always be supplemented by their empirical measures of dispersion like standard deviations and by confidence intervals to get an idea about their stability and variation, respectively. In any case, confidence intervals are more informative than the corresponding significance tests. Whereas the latter just lead to a binary decision, a confidence interval also allows the assessment of the uncertainty of an observed measure and of its practical relevance. Nevertheless, if p-values are used for exploratory purposes, they can be considered as an objective measure to judge the meaning of an observed association without declaring it as "statistically significant" or "non-significant." In conclusion, Rothman and Greenland (1998, p. 6) put it as follows: "The notion of statistical significance has come to pervade epidemiological thinking, as well as that of other disciplines. Unfortunately, statistical hypothesis testing is a mode of analysis that offers less insight into epidemiological data than alternative methods that emphasize estimation of interpretable measures."

Despite the justified condemnation of the uncritical use of statistical significance tests, they are widely used in the close-to-final step of an analysis to confirm or reject postulated research hypotheses (cf. next section). More sophisticated techniques such as multivariate regression models are applied in order to describe the functional relationship between exposures and outcomes (see chapters ▶Analysis of Continuous Covariates and Dose-Effect Analysis and ▶Regression Methods for Epidemiological Analysis of this handbook). Such techniques are an important tool to analyze complex data; but as it is the case with statistical tests, their application may lead to erroneous results if carried out without appropriately accounting for the epidemiological context. This, of course, holds for any statistical method. Its blind use may be misleading with serious consequences in practice. Therefore, each statistical analysis should be accompanied by sensitivity analyses (cf. chapter ▶Sensitivity Analysis and Bias Analysis of this handbook) and checks for model robustness. Graphical tools such as residual plots, for instance, to test for the appropriateness of a chosen statistical model should also routinely be used.

The final step concerns the adequate reporting of the results and their careful interpretation. The latter has to be done in light of background information and subject-matter knowledge about the investigated research area.

1.5.2 Statistical Thinking

According to the definitions quoted in Sect. 1.2.1, epidemiology deals with the distribution and determinants of health-related phenomena in *populations* as opposed to looking at *individual* subjects or cases. Studying distributions and their determinants in populations in a quantitative way is the very essence of statistics. In this sense, epidemiology requires statistical thinking (for an introduction to statistical inference, see chapter ▶Statistical Inference and to Bayesian methods see chapter ▶Bayesian Methods in Epidemiology of this handbook) in the context of health including the emphasis on causal analysis as described in chapter ▶Basic Concepts and the manifold applications to be found all over in this handbook. However, this conception of epidemiology started to permeate the field relatively late, and, at the beginning, often unconsciously.

The traditional separation of statistics into its descriptive and its inferential component has existed in epidemiology, too, until the two merged conceptually though not structurally. The *descriptive* approaches, introduced by scientists like Farr (see Sect. 1.3.1), continue in the form of *health statistics*, *health yearbooks*, and similar summary statistics by major hospitals, research organizations, and various health administrations like national ministries of health and the World Health Organization, often illustrated by graphics. The visual representation of the geographical distribution of diseases has recently taken an upsurge with the advent of *geographical information systems* (chapter ▶Geographical Epidemiology).

Forerunners of the use of *inferential* statistics in various parts of epidemiology are also mentioned in Sect. 1.3.1. Thus, in the area of *clinical trials*, the efficacy of citrus fruit to cure scurvy was established by purely statistical reasoning. In the realm of *causal factors* for diseases, the discovery of water contamination as a factor for cholera still relied on quite rudimentary statistical arguments, whereas the influence of the presence of a doctor at childbirth on maternal mortality was confirmed by a quantitative argument coming close to a modern test of significance. The basic statistical idea of *comparing frequencies* in populations with different levels of factors (or "determinants") was already present in all of these early investigations. The same is true for statistics in the domain of *diagnosis* where statistical thinking expresses itself by concepts like *sensitivity* or *specificity* of a medical test. It seems that this was only recently conceived of as a branch of epidemiology on par with the others, indispensable in particular for developing areas like computer-aided diagnosis or telediagnosis.

A landmark of statistical thinking in epidemiology was the elaboration of the theory of *hypothesis testing* by Neyman and Pearson. No self-respecting physician writes any more a paper on health in a population without testing some hypotheses on the significance level 5% or without giving a p-value. Most of these hypotheses are about the efficacy of a curative treatment or, to a lesser degree, the etiology of an ailment, but the efficacy of preventive treatments and diagnostic problems is also addressed.

However, the underlying statistical thinking was often deficient. Non-acceptance of the alternative hypothesis was frequently regarded as acceptance of the null hypothesis. The meaning of an arbitrarily chosen significance level or of a p-value was not understood, and in particular several simultaneous trials or trials on several hypotheses were not handled correctly by confusing the significance level of each part of the study with the overall significance level, that is, by not adjusting the single significance levels for multiple testing. Other statistical procedures that usually provide more useful insights like *confidence bounds* were neglected. Above all, *causal interpretations* were often not clear or outright wrong, and hence erroneous practical conclusions were drawn. A statistical result in the form of a hypothesis accepted either by a test or by a correlation coefficient far from 0 was regarded as final evidence and not as one element that should lead to further investigations, usually of a biological nature.

Current statistical thinking expresses itself mainly in the study of the effect of several factors on a health phenomenon, be it a causal effect in etiological research (chapter ▸Basic Concepts of this handbook), a curative or preventive effect in clinical or intervention trials, or the effect of a judgment, for example, a medical test. Such effects are expressed in quantitative, statistical terms. Here, relations between the action of several factors as described by the concepts of interaction and confounding play a prominent role (chapter ▸Confounding and Interaction of this handbook). The use of modern statistical ideas and tools has thus allowed for a conceptual and practical *unification* of the many parts of epidemiology. The *same* statistical models and methods of analysis (see chapters of Part III (Statistical Methods in Epidemiology) of the handbook) are being used in all of them.

1.5.3 Multivariate Analysis

An epidemiological study typically involves a huge number of variables to be collected from the study participants, which implies a high-dimensional dataset that has to be appropriately analyzed to extract the essential information. This curse of dimensionality becomes especially serious in genetic or molecular epidemiological studies. In such situations, statistical methods are called for to reduce the dimensionality of the data and to reveal the underlying association structure. Various multivariate techniques are at hand depending on the structure of the data and the research aim. They can roughly be divided into two main groups. The first group covers methods to structure the dataset without distinguishing response and explanatory variables, whereas the second group provides techniques to model and test for postulated dependencies. Although these multivariate techniques seem to be quite appealing at a first glance, they are not a statistical panacea. Their major drawback is that they cannot be easily understood by the investigator which typically leads to a less deep understanding of the data. This "black box" phenomenon also implies that the interpretation and communication of the results is not as straightforward as it is when just showing some well-known risk measures supplemented by frequency tables. In addition, the various techniques may not lead to the same result, while each of them may be compatible with the observed data. Thus, a final decision on the underlying data structure should not be made without critically reflecting the results in light of the epidemiological context and of additional subject-matter knowledge as well as compared with simpler statistical analyses such as stratified analyses – perhaps restricted to some key variables – that hopefully support the results obtained from the multivariate analysis.

Multivariate analyses with the aim to structure the dataset comprise factor analysis, cluster, and discriminatory analysis. Factor analysis tries to collapse a large number of observed variables into a smaller number of possibly unobservable, that is, latent variables, so-called factors, for example, in the development of scoring systems. These factors represent correlated subgroups of the original set. They serve in addition to estimate the fundamental dimensions underlying the observed dataset. Cluster analysis simply aims at detecting highly interrelated subgroups of the dataset, for example, in the routine surveillance of a disease. Having detected certain subgroups, their common characteristics might be helpful, for example, to identify risk factors, prevention strategies, or therapeutic concepts. This is in contrast to discriminant analysis, which pertains to a known number of groups and aims to assign a subject to one of these groups (populations) based on certain characteristics of this subject while minimizing the probability of misclassification. As an example, a patient with a diagnosis of myocardial infarction has to be assigned to one of the two groups, one consisting of long-term survivors of such an event and the other consisting of short-term survivors. For this purpose, the physician may measure his/her systolic and diastolic blood pressure, heart rate, and mean arterial pressure and assess his/her stroke index. With these medical parameters the physician will be able to predict whether or not the patient will die within a short period of time. A detailed discussion of these techniques is beyond the scope of this handbook. Instead, we refer to classical textbooks in this field such as Dillon and Goldstein (1984); Everitt and Dunn (2001), and Giri (2004).

However, in line with the idea of epidemiology, epidemiologists are often not so much interested in detecting a structure in the dataset rather than in explaining the occurrence of some health outcome depending on potentially explanatory variables. Here, it is hardly sufficient to investigate the influence of a single variable on disease risk as most diseases are the result of the complex interplay of many different exposure variables. Although it is very helpful to look first at simple stratified 2 × 2 tables to account for confounders, such techniques become impractical if a large number of variables have to be accounted for, especially if the study population is small. In such situations, techniques are needed that allow the examination of the effect of several variables simultaneously not only for adjustment but also for prediction purposes.

This is realized by regression models that offer a wide variety of methods to capture the functional relationship between outcomes and explanatory variables (see chapter ►Regression Methods for Epidemiological Analysis of this handbook). Models with more than one explanatory variable are usually referred to as multiple regression models, multivariable, or multivariate models where the latter might also involve more than one outcome. Using such techniques one needs to keep in mind that a statistical model rests on assumptions like normality, variance homogeneity, independence, and linearity that have all to be checked carefully in a given data situation. The validity of the model depends on these assumptions which might not be fulfilled by the data. Various models are therefore available from which an adequate one has to be selected. This choice is partly based on the research question and the a priori epidemiological knowledge about the relevant variables and their measurement. Depending on the scale, continuous or discrete, linear or logistic regressions might then be used for modeling purposes. Even more complex techniques such as generalized linear models can be applied where the functional relationship is no longer assumed to be linear or generalized estimation equations if a correlation structure has to be accounted for (see chapters ►Analysis of Continuous Covariates and Dose-Effect Analysis, ►Regression Methods for Epidemiological Analysis, and ►Generalized Estimating Equations of this handbook). Once the type of regression model is determined, one has to decide which and how many variables should be included in the model. If variables are strongly correlated with each other, only one of them should be included if justified from a subject-matter point of view. Many software packages offer automatic selection strategies such as forward or backward selection, which usually lead to different models that are all consistent with the data at hand. An additional problem may occur due to the fact that the type of regression model will have an impact on the variables to be selected and vice versa. The resulting model may also fail to recognize effect modification or may be affected by peculiarities of this particular dataset that are of no general relevance.

Further extensions of simple regression models are, for example, time-series models that allow for time-dependent variation and correlation, Cox-type models to be applied in survival analysis (see chapter ►Survival Analysis of this handbook), and the so-called graphical chain models which try to capture even more complex association structures. One of their features is that they allow for indirect influences by incorporating so-called intermediate variables that simultaneously serve as explanatory and response variables. The interested reader is referred to Lauritzen and Wermuth (1989), Wermuth and Lauritzen (1990), Whittaker (1990), Lauritzen (1996), and Cox and Wermuth (1996). In particular, directed acyclic graphs (DAGs) are not only used for data-driven modeling of complex association structures, but they have also become increasingly popular to postulate causal relationships, where the data-driven modeling aspect of DAGs can at best assist to explore causal dependence networks. This results from the fact that, so far, a single DAG cannot accomplish probabilistic and causal interpretations simultaneously (see chapter ▸Directed Acyclic Graphs of this handbook).

1.5.4 Handling of Data Problems

Data are the basic elements of epidemiological investigation and information. In the form of values of predictor variables, they represent levels of factors (risk factors and covariates), which are the *determinants* of health-related states or events in the sense of the definition of epidemiology quoted in Sect. 1.2.1. As values of *response* (outcome) variables, they describe the health-related phenomena themselves. Measuring these values accurately is obviously fundamental for the conclusions to be drawn from an epidemiological study. The practical problems that arise when trying to do this are outlined in chapters ▸Design and Planning of Epidemiological Studies, ▸Quality Control and Good Epidemiological Practice, ▸Epidemiological Field Work in Population-Based Studies, and ▸Exposure Assessment of this handbook. However, even when taking great care and applying a rigorous quality control, some of the recorded data may still be erroneous and others may be missing. The question of how to handle these problems is the subject of chapters ▸Measurement Error and ▸Missing Data of this handbook.

Intuitively, it is clear that in both cases the approach to be taken depends on the particular situation, more precisely, on the type of *additional information* that may be available. We use this information either to correct or to supplement certain data individually or to correct the final results of the study.

Sometimes a naïve approach looks sensible for replacing a missing value or correcting a measurement error. Thus, if we know that the data at hand represent the size of a tumor in consecutive months, we may be tempted to replace a missing or obviously out-of-range value by an interpolated one. When monitoring maternal mortality in a developing country by studies carried out routinely on the basis of death registers, we may multiply the figures obtained by a factor that reflects the fact that many deaths in childbed are not recorded in these registers. This factor may be estimated beforehand by special studies where all such deaths were searched for, for example, by visits to the homes of diseased women and retrospective interviews. Even with such simple procedures, though, it is difficult to assess their impact on the *statistical quality* of a study, be it the power of a test or the width of a confidence interval.

The basic idea underlying the rigorous handling of measurement errors is to represent the *true* predictor variables that cannot be measured exactly via the so-called *surrogate* predictor variables that can be measured error free. The way a surrogate and a predictor are assumed to be related and the corresponding distributional assumptions form the so-called *measurement error model*. Several types of such models have been suggested and explored, the goal always being to get an idea about the magnitude of the effect on the statistical quality of the study if we correct the final results depending on the postulated model.

The general ideas that underlie methods for dealing with missing values are similar although the technical details are quite different. These methods are mainly based on jointly modeling the predictor and response variables and the missing value mechanism. This mechanism may or may not allow for imputation of a missing value.

1.5.5 Meta-Analysis

The use of meta-analyses to synthesize the evidence from epidemiological studies has become more and more popular. They can be considered as the quantitative parts of systematic reviews. Usually the main objective of a meta-analysis is the statistical combination of results from several studies that individually are not powerful enough to demonstrate a presumed effect. However, whereas it is always reasonable to review the literature and the published results on a certain topic systematically, the statistical combination of results from separate epidemiological studies may yield misleading results. Purely observational studies are in contrast to randomized clinical trials where differences in treatment effects between studies can mainly be attributed to random variation. Observational studies, however, may lead to different estimates of the same effect that can no longer be explained by chance alone, but that may be due to confounding and bias potentially inherent in each of them. Thus, the calculation of a combined measure of association based on heterogeneous estimates arising from different studies may lead to a biased estimate with spurious precision. Although it is possible to allow for heterogeneity in the statistical analysis by so-called random-effects models, their interpretation is often difficult. Inspecting the sources of heterogeneity and trying to explain it would therefore be a more sensible approach in most instances.

Nevertheless, a meta-analysis may be a reasonable way to integrate findings from different studies and to reveal an overall trend in the data. A meta-analysis of several studies to obtain an overall estimate of association, for instance, can be performed by pooling the original data or by calculating a combined measure of association based on the study-specific estimates. In both cases, it is important to retain the study as the unit of analysis. Ignoring this fact would lead to biased results since the between- and within study variations and the different sample sizes would then not be adequately accounted for in the statistical analysis.

Since the probable first application of formal methods to pool several studies by Pearson (1904), numerous sophisticated statistical methods have been developed that are reviewed in chapter ▸Meta-Analysis in Epidemiology of this handbook.

1.6 Research Areas

Epidemiology pursues three major objectives: (1) to describe the spectrum of diseases and their determinants, (2) to identify the causal factors of diseases, and (3) to apply this knowledge to prevention and public health practice.

1.6.1 Description of the Spectrum of Diseases

Describing the distribution of disease is an integral part of the planning and evaluation of health-care services. Often, information on possible exposures and disease outcomes has not been gathered with any specific hypothesis in mind but stems from routinely collected data. These descriptions serve two main purposes. First, they help in generating etiological hypotheses that may be investigated in detail by analytical studies. Second, descriptive data form the basis of health reports that provide important information for the planning of health systems, for example, by estimating the prevalence of diseases and by projecting temporal trends. The approaches used in descriptive epidemiology are presented in chapter ▸Descriptive Studies of this handbook.

Complementary descriptive information relates to the revelation of the natural history of diseases – one of the subjects of clinical epidemiology – that helps to improve diagnostic accuracy and therapeutic procedures in the clinical setting. The understanding of a disease process and its intermediate stages also gives important input for the definition of outcome variables, be it disease outcomes that are used in classical epidemiology or precursors of disease and preclinical stages that are relevant for screening or in molecular epidemiology studies.

1.6.2 Identification of Causes of Disease

A major part of current research in epidemiology is tied to the general methodological issues summarized in Sects. 1.4 and 1.5. These concern *any* kind of exposures (risk factors) and *any* kind of outcomes (health defects). After having described the basic ideas and the main procedures, the emphasis is now on more specific questions determined by a particular type of exposure (see chapters in Part IV (Exposure-Oriented Epidemiology) of this handbook) or a special kind of outcome (see chapters in Part V (Outcome-Oriented Epidemiology) of this handbook) or both. This distinction, however, is somewhat arbitrary, as some research areas are not easily assigned to either category as, for example, molecular epidemiology, clinical epidemiology, and infectious disease epidemiology.

1.6.2.1 Exposure-Oriented Research

The search for factors that cause a disease is a central feature of epidemiology. This is nicely illustrated by the famous investigation into the causes of cholera by John Snow, who identified the association of poor social and hygienic conditions, especially of the supply with contaminated water, with the disease which provided the basis for preventive action. Since that time, the investigation of hygienic conditions has been diversified by examining infectious agents, nutrition, pharmaceuticals, social conditions, or lifestyle factors as well as physical and chemical agents in the environment or at the workplace. A peculiarity is the investigation of genetic determinants by themselves and their interaction with the extraneous exposures mentioned above that requires special statistical methods for their analysis as outlined in chapter ▸Statistical Methods in Genetic Epidemiology of this handbook.

Nutrition belongs to the most frequently studied exposures and may serve as a model for both the methodological problems of exposure-oriented research and its potential for public health (cf. chapter ▸Nutritional Epidemiology of this handbook). There are only few health outcomes for which nutrition does not play a role in causation. A solid evidence base is required to guide action aiming at disease prevention and improvement of public health. Poor nutrition has direct effects on growth and normal development, as well as on the process of healthy aging. For example, 25% of cancer burden were estimated to be attributable to nutrition and obesity (Doll and Peto 2005). The effect of poor diet on chronic diseases is complex, for example, the role of micronutrients in maintaining optimal cell function and reducing the risk of cancer and cardiovascular disease. The accurate assessment of nutrient intake is a major challenge in observational studies, and available instruments are prone to bias and measurement error. Moreover, foods contain more than nutrients, and the way foods are prepared may enhance or reduce their harmful or beneficial effects on health. Because diet and behavior are complex and interrelated, it is important, both in the design and the interpretation of studies, to understand how this complexity may affect the results.

Accurate measurement of exposure is a particular problem in retrospective studies that exposure-oriented epidemiology is faced with. The use of biological markers of exposure and early effect has been proposed to reduce exposure misclassification. In a few cases, biomarker-based studies have led to important advances, as, for example, illustrated by the assessment of exposure to aflatoxins, enhanced sensitivity and specificity of assessment of past viral infections, and detection of protein and DNA adducts in workers exposed to reactive chemicals such as ethylene oxide. In other cases, however, several initially promising findings have not been confirmed by more sophisticated investigations. They include in particular the search for susceptibility to environmental carcinogens by looking at polymorphism for metabolic enzymes. The new opportunities offered by biomarkers to overcome some of the limitations of traditional approaches in epidemiology need to be assessed systematically. The measurement of biomarkers should be quality-controlled and validated. Sources of bias and confounding in molecular epidemiology studies have to be assessed with the same scruting as in other types of epidemiological studies.

Modern molecular techniques have made it possible to investigate exposure to genetic factors in the development or the course of diseases on a large scale (cf. chapter ▸Molecular Epidemiology of this handbook). Many diseases aggregate in families. Although some of the aggregation can be explained by shared risk factors among family members, it is plausible that a true genetic component exists for most human cancers and for the susceptibility to many infectious diseases. The knowledge of low-penetrance genes responsible for such susceptibility is still very limited, although research has focused on genes encoding for metabolic enzymes, DNA repair, cell-cycle control, and hormone receptors. In many studies only indirect evidence can be used since the suspected disease-related gene (candidate gene) is not directly observable. To locate or to identify susceptibility genes, genetic markers are used either in a so-called whole genome scan or in the investigation of candidate genes. The latter can be performed by linkage studies, where the common segregation of a marker and a disease is investigated in pedigrees; and by association studies, where it is investigated whether certain marker alleles of affected individuals are more or less frequent than in a randomly selected individual from the population. Both, population-based and family designs are complementary and play a central role in genetic epidemiological studies. In the case of low-penetrance genes, association studies have been successful in identifying genetic susceptibility factors. Given the lack of dependence of genetic markers from time of disease development, the case-control approach is particularly suitable for this type of investigation because their assessment is not prone to recall bias. More pronounced than in classical epidemiology, the three main complications in genetic epidemiology are dependencies, use of indirect evidence, and complex datasets.

1.6.2.2 Outcome-Oriented Research

Epidemiology in industrialized countries is nowadays dominated by research on chronic diseases, among them cardiovascular diseases, cancer, and metabolic, respiratory, digestive, psychiatric, and musculoskeletal disorders. Their epidemiology – especially the one of cancer – is characterized more than any other outcome-defined epidemiology by the abundance of observational studies to identify risk factors of all kind.

Cardiovascular diseases (cf. chapter ▸Cardiovascular Health and Disease of this handbook) have a multifactorial etiology and confounding effects are especially intriguing. For example, clustering of coronary heart diseases in families could be due both to genetic factors and to common dietary habits. High blood pressure and obesity play both the role of an outcome variable and of a risk factor. A typical feature of the epidemiology of cardiovascular diseases is marked by the decline of morbidity and mortality in some areas and population groups whose causes are manifold, including, for example, control of blood pressure and blood cholesterol as well as improved treatment regimens.

In many respects cancer epidemiology exemplifies the strengths and the weaknesses of the discipline at large. Although it is a relatively young discipline, it has been the key tool to demonstrate the causal role of important cancer risk factors, like smoking, human papilloma virus infections in cervical cancer, solar radiation in skin cancer, and obesity in many neoplasms (cf. chapter ▶Cancer Epidemiology of this handbook). Cancer epidemiology is an area in which innovative methodological approaches are developed as illustrated by the increasing use of biological and genetic markers pertaining to causal factors and early outcomes.

By comparison, the epidemiology of musculoskeletal disorders is still less developed. Already the definition of the various disorders and the distinction between them are still subject to debate. Case ascertainment is often tricky. In spite of the high prevalence of, for example, back pain or osteoarthritis and their enormous negative impact on quality of life, mortality caused by them is significantly lower than that by cancer or cardiovascular diseases. Even simple estimates of prevalence leave wide margins. Regarding established risk factors, we find, for instance, for osteoarthritis, and depending on its location, genetic factors, sex, obesity, heavy physical workload, and estrogen use. Not much more seems to be known although certain nutritional factors have been mentioned like red meat and alcohol (cf. chapter ▶Musculoskeletal Disorders of this handbook).

The investigation of infectious diseases is the most important historical root of epidemiology and is still of primary importance in developing countries. If a person suffers from a particular outcome like tuberculosis, the exposure "infection by the relevant microorganism," that is, by mycobacterium tuberculosis, must have been present by the very definition of the disease. However, it is not a sufficient condition for overt disease, and many analytical studies examine the influence of cofactors like social conditions, nutrition, and comorbidities regarded as risk factors for opportunistic infections. Purely descriptive health statistics, too, play a very important role in controlling infectious diseases. Related activities are general *epidemic surveillance*, *outbreak studies* by tracing possible carriers, and the search for infectious sources like salmonella as sources of food poisoning or the various origins of *nosocomial* illness. The most specific features of the epidemiology of infectious diseases are *mathematical modeling* and *prevention by immunization*. Modeling has to be understood in the sense of population dynamics. What is being modeled is typically the temporal development of the incidence or prevalence of the disease in question. The model, be it deterministic or stochastic, describes the mechanism of the infection. It depends on specific parameters like contact frequencies between infected and susceptible subjects and curing rates. It is interesting to note that with the discovery of the infectious cycle of malaria, Ronald Ross, also designed and analyzed a mathematical model for it that led him to conceive the *threshold principle* (Bailey 1975; Diekmann and Heesterbeek 2000). The prevention of infectious diseases can in principle be done in three ways: by acting on cofactors of the type mentioned above; by interfering with the infectious process via hygiene, separation of susceptible persons from carriers or vectors, and elimination of vectors; or by raising the immunity of susceptible people by various measures like preventive drug treatment, the main method of immunization being a vaccination, though. The effect of a vaccination in a population can be modeled in its turn, which leads in particular to the basic epidemiological concept of *herd immunity* (chapter ▶Infectious Disease Epidemiology of this handbook).

References

Ackerknecht EH (1953) Rudolf Virchow: Doctor – statesman – anthropologist. University of Wisconsin Press, Madison

Ackerknecht EH (1967) Medicine at the Paris hospital, 1794–1848. Johns Hopkins, Baltimore

Ahrens W (1999) Retrospective assessment of occupational exposure in case-control studies. ecomed Verlagsgesellschaft, Landsberg

Altman DG, Bland JM (1995) Statistics notes: absence of evidence is not evidence of absence. Br Med J 311:485

Bailey NTJ (1975) The mathematical theory of infectious diseases and its applications. Griffin, London

Bernard C (1865) Introduction à l'étude de la médecine expérimentale. J.B. Baillière et Fils, Paris

Bernoulli D (1766) Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, & des avantages de l'inoculation pour la prévenir. Mém Math Phys Acad Roy Sci, Paris

Boffetta P, Agudo A, Ahrens W, Benhamou E, Benhamou S, Darby SC, Ferro G, Fortes C, Gonzalez CA, Jöckel KH, Krauss M, Kreienbrock L, Kreuzer M, Mendes A, Merletti F, Nyberg F, Pershagen G, Pohlabeln H, Riboli E, Schmid G, Simonato L, Tredaniel J, Whitley E, Wichmann HE, Saracci R (1998) Multicenter case-control study of exposure to environmental tobacco smoke and lung cancer in Europe. J Natl Cancer Inst 90:1440–1450

Buell P, Dunn JE (1965) Cancer mortality among Japanese Issei and Nisei of California. Cancer 18:656–664

Bulloch W (1938) The history of bacteriology. Oxford University Press, London

Chan-Yeung M, Dimich-Ward H (2003) Respiratory health effects of exposure to environmental tobacco smoke. Respirology 8:131–139

Chen WY, Colditz GA, Rosner B, Hankinson SE, Hunter DJ, Manson JE, Stampfer MJ, Willett WC, Speizer FE (2002) Use of postmenopausal hormones, alcohol, and risk for invasive breast cancer. Ann Intern Med 137:798–804

Cox D, Wermuth N (1996) Multivariate dependencies – models, analysis and interpretation. Chapman and Hall, London

Coxe JR (1846) The writings of Hippocrates and Galen. Lindsay and Blackiston, Philadelpia

Dawber TR (1980) The Framingham study: the epidemiology of atherosclerotic disease. Harvard University Press, Cambridge, Mass

Dawber TR, Meadors GF, Moore FE Jr (1951) Epidemiological approaches to heart disease: the Framingham study. Am J Public Health 41:279–286

Detels R (2002) Epidemiology: the foundation of public health. In: Detels R, McEwen J, Beaglehole R, Tanaka H (eds) Oxford textbook of public health. Vol 2: The methods of public health, 4th edn. Oxford University Press, Oxford/New York, pp 485–491

Detels R, Breslow L (2002) Current scope and concerns in public health. In: Detels R, McEwen J, Beaglehole R, Tanaka H (eds) Oxford textbook of public health. Vol 1: The scope of public health, 4th edn. Oxford University Press, Oxford/New York, pp 3–20

Diekmann O, Heesterbeek JAP (2000) Mathematical epidemiology of infectious diseases. Wiley, Chichester

Dillon W, Goldstein M (1984) Multivariate analysis, methods and applications. Wiley, New York

Doll R, Hill AB (1950) Smoking and carcinoma of the lung: preliminary report. Br Med J 2: 739–748

Doll R, Hill AB (1954) The mortality of doctors in relation to their smoking habits: a preliminary report. Br Med J 1:1451–1455

Doll R, Hill AB (1964) Mortality in relation to smoking: ten years' observation of British doctors. Br Med J 1:1399–1410, 1460–1467

Doll R, Peto R (1978) Cigarette smoking and bronchial carcinoma: dose and time relationships among regular smokers and lifelong non-smokers. J Epidemiol Community Health 32: 303–313

Doll R, Peto R (2005) Epidemiology of cancer. In: Warrell DA, Cox TM, Firth JD (eds) Oxford textbook of medicine, 4th edn. Oxford University Press, Oxford, pp 193–218

Doll R, Wald NJ (1994) Interpretation of negative epidemiological evidence for carcinogenicity. IARC Scientific Publications, vol 65. International Agency for Research on Cancer (IARC), Lyon

Everitt B, Dunn G (2001) Applied multivariate data analysis, 2nd edn. Edward Arnold, London

Farr W (1975) Vital statistics: a memorial volume of selections from the reports and writings of William Farr. New York Academy of Sciences. Scarecrow Press, Metuchen

Fracastoro G (1521) De contagione et contagiosis morbis et eorum curatione. English translation by Wright WC (1930). G P Putnam's Sons, New York

Galen (1951) De sanitate tuenda. English translation by Green R. C C Thomas, Springfield

Giri NC (2004) Multivariate statistical analysis, 2nd edn. Marcel Dekker, New York

Gordon A (1795) A treatise on the epidemic puerperal fever of Aberdeen. Printed for G.G. and J. Robinson, London

Graunt J (1662) Natural and political observations mentioned in a following index, and made upon the bills of mortality ... with reference to the government, religion, trade, growth, ayre, diseases, and the several changes of the said city .... Martin, Allestry, and Dicas, London

Greenwood M (1932) Epidemiology, historical and experimental. Johns Hopkins, Baltimore

Halley E (1693) An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the city of Breslau; with an attempt to ascertain the price of annuities upon lives. Philos Trans R Soc London 17:596–610, 654–656

Harstad I, Heldal E, Steinshamn SL, Garåsen H, Jacobsen GW (2009) Tuberculosis screening and follow-up of asylum seekers in Norway: a cohort study. BMC Public Health 9:141

Hennekens CH, Buring JE (1987) Epidemiology in medicine. Little, Brown, Boston/Toronto

Hill AB (1962) Statistical methods in clinical and preventive medicine. Livingstone, Edinburgh

Hill AB (1965) The environment and disease: association or causation? Proc R Soc Med 58: 295–300

Hippocrates (400 BC, approximately) On airs, waters, and place. There exist several editions in English, among others in: (1846) Coxe, see above; (1939) The genuine works of Hippocrates. Williams and Wilkins, Baltimore; (1983) Lloyd (ed) Hippocratic writings. Penguin Classics, Harmondsworth. http://classics.mit.edu/Hippocrates/airwatpl.html. Accessed 18 April 2013

Holmes OW (1842–1843) The contagiousness of puerperal fever. N Engl Q J Med Surg 1:503–540

IARC (2004) Tobacco smoke and involuntary smoking. IARC Monographs, Vol 83. International Agency for Research on Cancer (IARC), Lyon

Jenner E (1798) An inquiry into the causes and effects of the variolae vaccinae, a disease known by the name of Cow Pox. Published by the author, London

Keys A (1980) Seven Countries: a multivariate analysis of death and coronary heart disease. Harvard University Press, Cambridge

Khoury MJ, Beaty TH, Cohen BH (1993) Fundamentals of genetic epidemiology. Oxford University Press, New York

Kromhout D, Menotti A, Bloemberg B, Aravanis C, Blackburn H, Buzina R, Dontas AS, Fidanza F, Giampaoli S, Jansen A, Karvonen M, Katan M, Nissinen A, Nedeljkovic S, Pekkanen J, Pekkarinen M, Punsar S, Räsänen L, Simic B, Toshima H (1995) Dietary saturated and trans fatty acids and cholesterol and 25-year mortality from coronary heart disease: the Seven Countries study. Prev Med 24:308–315

Kromhout H (1994) From eyeballing to statistical modelling. Methods for assessment of occupational exposure. Landbouwuniversiteit de Wageningen, pp 1–210

Last JM (ed) (2001) A dictionary of epidemiology, 4th edn. Oxford University Press, New York/Oxford/Toronto

Lauritzen SL (1996) Graphical models. Oxford University Press, Oxford

Lauritzen SL, Wermuth N (1989) Graphical models for associations between variables, some of which are qualitative and some quantitative. Ann Stat 17:31–57

Lee PN, Forey BA (1996) Misclassification of smoking habits as a source of bias in the study of environmental tobacco smoke and lung cancer. Stat Med 15:581–605

Lind J (1753) A treatise on the scurvy. Sands, Murray, and Cochran, Edinburgh

Lind J (1771) An essay on diseases incidental to Europeans in hot climates with the method of preventing their fatal consequences, 2nd edn. T Becket and PA De Hondt, London

Lipnick RJ, Buring JE, Hennekens CH, Rosner B, Willett W, Bain C, Stampfer MJ, Colditz GA, Peto R, Speizer FE (1986) Oral contraceptives and breast cancer. A prospective cohort study. JAMA 255:58–61

Louis PCA (1835) Recherches sur les effets de la saignée dans quelques maladies inflammatoires et sur l'action de l'émétique et les vésicatoires dans la pneumonie. Baillière, Paris

MacMahon B, Pugh TF (1970) Epidemiology: principles and methods. Little, Brown, Boston

Marshall H, Tulloch AM (1838) Statistical report on the sickness, mortality and invaliding in The West Indies. Prepared from the records of the Army Medical Department and War Office returns. Clowes & Sons, London

Mausner JS, Bahn AK (1974) Epidemiology: an introductory text. W.B. Saunders, Philadelphia/London/Toronto

Nieuwenhuijsen JM (ed) (2003) Exposure assessment in occupational and environmental epidemiology. Oxford University Press, Oxford

Olsen J, Ahrens W, Björner J, Grönvold M, Jöckel K-H, Keiding L, Lauritsen JM, Levy-Desroches S, Manderson L, Olesen F (1998) Epidemiology deserves better questionnaires. Int J Epidemiol 27(6):935

Pagel W (1982) Paracelsus. An introduction to philosophical medicine in the era of the Renaissance, 2nd edn. Karger, Basel

Pearson K (1904) Report on certain enteric fever inoculation statistics. Br Med J 3:1243–1246

Porta M (ed) (2008) Dictionary of epidemiology, 5th edn. Oxford University Press, New York

Pott P (1775) Chirurgical observations relative to the cataract, the polypus of the nose, the cancer of the scrotum, the different kinds of ruptures, and the mortification of the toes and feet. Hawes, Clarke and Collins, London

Ramazzini B (1713) De morbis artificum diatriba. Baptistam Conzattum, Padua

Riboli E, Preston-Martin S, Saracci R, Haley NJ, Trichopoulos D, Becher H, Burch JD, Fontham ETH, Gao Y-T, Jindal SK, Koo LC, Marchand LL, Segnan N, Shimizu H, Stanta G, Wu-Williams AH, Zatonski W (1990) Exposure of nonsmoking women to environmental tobacco smoke: a 10-country collaborative study. Cancer Causes Control 1:243–252

Rose G, Shipley M (1986) Plasma cholesterol concentration and death from coronary heart disease: 10 year results of the Whitehall study. Br Med J 293:306–307

Rothman KJ, Greenland S (1998) Modern epidemiology, 2nd edn. Lippincott-Raven, Philadelphia

Rothman N, Hainaut P, Schulte P, Smith M, Boffetta P, Perera F (2012) Molecular epidemiology: principles and practices. IARC Scientific Publication, No 163. International Agency for Research on Cancer, Lyon

Rupke N (ed) (2000) Medical geography in historical perspective. Medical History, Supplement No. 20. The Wellcome Trust Centre for the History of Medicine at UCL, London

Schairer E, Schöniger E (1943) Lungenkrebs und Tabakverbrauch. Z Krebsforsch 54:261–269 (English translation: Schairer E, Schöniger E (2001) Lung cancer and tobacco consumption. Int J Epidemiol 30:24–27)

Schulte P, Perera FP (eds) (1993) Molecular epidemiology: principles and practices. Academic Press, New York

Semmelweis IP (1861) Die Aetiologie, der Begriff and die Prophylaxis des Kindbettfiebers. Hartleben, Pest-Wien-Leipzig

Siegel RE (1968) Galen's system of physiology and medicine. Karger, Basel

Simpson JY (1868–1869, 1869–1870) Our existing system of hospitalism and its effects. Edinb Mon J Med Sci 14:816–830, 1084–1115; 15:523–532

Snow J (1855) On the mode of communication of cholera, 2nd edn. Churchill, London

Stampfer MJ, Willett WC, Colditz GA, Rosner B, Speizer FE, Hennekens CH (1985) A prospective study of postmenopausal estrogen therapy and coronary heart disease. N Engl J Med 331:1044–1049

Taubes G (1995) Epidemiology faces its limits. Science 269:164–169

Tenon JR (1788) Mémoires sur les hôpitaux de Paris. Pierres, Paris

Toniolo P, Boffetta P, Shuker D, Rothman N, Hulka B, Pearce N (1997) Application of biomarkers to cancer epidemiology. IARC Scientific Publications, No. 142. International Agency for Research on Cancer, Lyon

Truett J, Cornfield J, Kannel E (1967) Multivariate analysis of the risk of coronary heart disease in Framingham. J Chronic Dis 20:511–524

Wells AJ, English PB, Posner SF, Wagenknecht LE, Perez-Stable EJ (1998) Misclassification rates for current smokers misclassified as smokers. Am J Public Health 88:1503–1509

Wermuth N, Lauritzen SL (1990) On substantive research hypotheses, conditional independence, graphs and graphical chain models. J R Stat Soc Ser B 52:21–50

White E, Armstrong BK, Saracci R (2008) Principles of exposure measurement in epidemiology: collecting, evaluating and improving measures of disease risk factors, 2nd edn. Oxford University Press, Oxford/New York

White KL, Henderson MM (1976) Epidemiology as a fundamental science. Its uses in health services planning, administration, and evaluation. Oxford University Press, New York

Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley, Chichester

Willett WC, Stampfer MJ, Colditz GA, Rosner BA, Hennekens CH, Speizer FE (1987) Dietary fat and the risk of breast cancer. N Engl J Med 316:22–28

World Medical Association (2000) The revised Declaration of Helsinki. Interpreting and implementing ethical principles in biomedical research. Edinburgh: 52nd WMA General Assembly. JAMA 284:3043–3045

History of Epidemiological Methods and Concepts

2

Alfredo Morabia

Contents

A. Morabia (✉)
Center for the Biology of Natural Systems, Queens College City University of New York,
New York, Flushing, NY, USA

Department of Epidemiology, Mailman School of Public Health, Columbia University,
New York, NY, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_52,

2.1 Introduction

Epidemiology emerged as a scientific discipline in the seventeenth century when the conditions became ripe for collecting, analyzing, and interpreting population data. Population comparative studies were performed in the eighteenth century. Since then, a body of methods and concepts needed to perform population studies has been developed, refined, and theorized. It constitutes the genuine core of the scientific discipline known today as epidemiology. I identify four phases in the evolution of these epidemiological methods and concepts. During the preformal phase (seventeenth to nineteenth century), there was no theory guiding the implementation of population studies. During the early phase (1900–1945), study designs were segregated, ways of dealing with non-comparability were developed and applied, epidemiology became an academic discipline, and epidemiological textbooks were published. In the third phase, classic epidemiology (1945–1980), study designs were refined, and the measurement of effects, confounding effects, interactions, and biases were formalized. The latest phase, modern epidemiology (since 1980), has been characterized by an integration of methods and concepts into a theoretical corpus that comprises study designs, measures of event occurrences and effects, confounding effects, interactions, biases, and causal inferences. There are indications that a new phase is dawning, characterized by refinements in conceptual domains such as the assessment of confounding effects, biases, and of complex interactions.

2.1.1 Note About Terminology and References

I use risk, rate, and odds, followed by “ratio” or “difference,” as defined by Elandt-Johnson (1975), to describe measures of effect. Two measures of disease occurrence are either divided (ratio) or subtracted (difference). The usage of these terms has varied substantially across time and place (Zhang et al. 2004). The resulting dilemma for the historian is between using the terms as found in the historical documents and using the modern terminology throughout. I used to let the terminology vary across historical periods in order to be consistent with the terms used in the historical documents I was describing until I realized that, besides those with some advanced epidemiological training, most students were confused by the multiplicity of terms used to define the same concept. I now prefer to stick to the modern terminology. For example, I don’t use the terms retrospective and prospective studies to characterize, respectively, case-control and cohort studies even though the former terms have been used for decades before being abandoned (Vandenbroucke 1991; Doll 1991). But I do qualify as “cohort” a German study published in 1913, that is, about half a century before the term was proposed as a synonym of prospective study. Substituting the ancient with the modern terminology may sometimes make the text sound anachronistic, but it also highlights the continuity in the evolution of epidemiological methods and concepts over the last three centuries.

Years of birth and death of scientists and philosophers, whose names are mentioned in this chapter, were only provided when both dates were potentially available.

For research and teaching purposes only, most papers cited in this article can be downloaded from the People's Epidemiology Library (www.epidemiology.ch/history/PeopleEpidemiologyLibrary.html).

2.2 A Historical Definition of Epidemiology

Epidemiology is a combination of Greek terms meaning, literally, the science (logos) of what falls upon (epi) the people (demos), or the science of epidemics. It was originally baptized as such, to characterize the study of epidemics of mostly infectious diseases (Villalba 1803). The name also implies that the need for epidemiology started with the first epidemics. When did the first epidemics occur?

2.2.1 Genesis and Evolution of Epidemics in Agrarian Societies

There are good reasons to believe that epidemics were not a major issue in foraging societies, which subsisted by hunting, gathering, and fishing and did not produce food. They consisted of small, mobile bands, usually less than 50 people, without domesticated animals (except for dogs in the latest period before the Neolithic). No infectious disease could become recurrent in these small, nomadic clans because a contagious and lethal disease would either wipe out the clan or be interrupted when the clan moved away from the outbreak, leaving its sick and dead behind. Data on prehistoric and contemporary foragers support the idea that no recurring epidemics existed before 10,000 BCE: a convenient transition date between foraging and agrarian societies.

Two conditions were needed for recurring epidemics to occur (McNeill 1977). For one, large domesticated animals living closely enough with humans had to provide an opportunity for some animal parasites, such as the agents of tuberculosis, measles, flu, smallpox, and so on, to adapt to the human organism. Secondly, a population pool of at least 500,000 people was necessary for a parasite to permanently subsist within a population and cyclically emerge when the pool of susceptible subjects is reconstituted. Early agrarian civilizations were therefore most likely free of epidemics since these two conditions were supposedly first met in Sumer, Mesopotamia, around 2000 BCE. The plagues that afflicted the Egyptians in the Book of Exodus may have occurred between 1000 BCE and 500 BCE. It is generally accepted that since then, epidemics of new (i.e., neolithic or later) infectious diseases disseminated throughout the globe until all the population basins were connected. At that stage, local epidemics evolved into world pandemics (McNeill 1977).

Data from China between 243 BCE and 1911 CE provide a glimpse into the evolution of the intensity of outbreaks during the past 2,000 years (Morabia 2009).

In China, the frequency of outbreaks has been proportional to the size of the population. The burden of epidemics increased exponentially after 1100 and continued unabated until the nineteenth century (Morabia 2009). It is probable that the evolution of epidemics was similar in Europe and in China, which were the two most populated areas in the Old World. Except maybe for the huge death toll of the Black Death epidemic during the fourteenth century, which affected European populations but maybe not China, there are indications that epidemics also grew in intensity between 1,000 and 1,800 CE in European countries, such as Britain (Creighton 1894).

Thus, historical evidence suggests that pre-epidemiology medical systems across continents failed, by and large, to understand and control the evolution of infectious diseases. Why?

2.2.2 Why Pre-Epidemiology Medical Systems Failed to Understand Epidemics

For about 4,000 years, since the emergence of medicine as a profession (Porter 1997), premodern doctors have had the ambitious aim of understanding the complex and unique set of causes responsible for the occurrence of *each specific case* of disease. They had a holistic vision of health and disease based on the belief that the body and the universe coexisted in precarious equilibrium. A disequilibrium meant disease. Factors that could precipitate disease were so numerous that each case had its own set of causes, and no two cases were alike. Except when it came to describe qualitatively the frequency of symptoms and signs among the sick, doctors focused exclusively on the individual patient and therefore could not view – and were probably not interested in – the pattern of disease occurring at the population level. Their clinical practice confronted them with the individual consequences of epidemics but not with their population effects. They were not conceptually equipped to study the causes of diseases and to assess the efficacy of treatments. From this point of view, Hippocratic medicine is a form of ancient holistic medicine.

The idea that each individual case is different subsisted until well into the nineteenth century. When cholera broke out in New York in 1832, the apothecary Horace Bartley (dates unknown) described each case admitted to a local hospital separately, without attempting to group them in any way. Doctors prescribed different treatments to each patient, hoping that one would work. But this trial-and-error approach could not succeed (Bartley 1832). Had they grouped and compared the patients by treatment received, they would have found that camphor, peppermint, mercury, and sulfur worked half of the time given that the lethality of cholera is about 50%. In reality, not a single one of these treatments was effective.

Prescribing a different treatment to each patient was compatible with the dominant theory of the early nineteenth century regarding disease causation. It was believed that diseases were due to some form of air pollution. The putrefaction of organic matters, such as human excrements, garbage, and dead animal bodies, which abounded in the streets of poor and working class neighborhoods, was thought to release toxic gases, sickening those who inhaled them. But the miasma theory required a second condition to explain why not everyone was being affected: miasma produced disease solely among people who were predisposed to it because of their constitution or because, in their individual context, body fluids were in disequilibrium. In other words, each case of disease was unique and needed to be treated in a specific way. Most people in medicine and public health, including both doctors and social reformers, largely adhered to this multifactorial theory, which was in line with the holistic beliefs of the past. It was therefore not a surprise that the late Hippocratic concept of miasma subsisted for over 2,500 years.

In the nineteenth century, the success of the miasma theory was based on its apparent ability to guide public health action. It implied that reducing poverty, cleaning streets, improving housing, and draining dirt and excrement by creating a proper sewage system would prevent disease. And it did. Cleaning lice prevented typhus; exterminating rats prevented plague; avoiding fecal pollution prevented typhoid and cholera. The apparent successes of the miasma theory blinded its supporters from its flaws, which were not discovered until scientists tested their hypotheses using population comparative studies.

2.3 Epidemics Since Epidemiology

2.3.1 Plague Surveillance

The first known analysis of population data to describe patterns of epidemics focused on the plague, which, since the fourteenth century, was a recurring scourge in most European countries. No one knew what caused it. Religious and superstitious reasons were often invoked. Even though its contagious nature was highly debated among scholars, people behaved as if it was: authorities sealed to death the sick and their families in their houses, and the upper class had formed a safe habit of fleeing to the countryside when the disease was lurking in their town.

In 1662, John Graunt (1620–1674) published an analysis of the trends in mortality from the London plague, based on the Bills of Mortality, which were the ancestors of our current death certificates. Graunt observed that there was regularity and predictability in the trends of deaths from causes other than plague. Plague, however, made “sudden jumps” in frequency, which was more consistent with an environmental origin than with a disorder originating from the “constitution of human bodies” (Graunt 1662). Graunt did not mention astrologic or divine factors. The last outbreak of plague in London occurred in 1665, shortly after Graunt’s publication. The exact reasons for the interruption of the plague cycle are unknown, but it is reasonable to believe that once population evidence clearly showed that plague was caused by an environmental factor that cyclically came and went, authorities began to enforce ship quarantines and cordon sanitaires on Europe eastern borders more rigorously and more effectively than in the past.

2.3.2 Nineteenth-Century Population Sciences

Graunt was a pioneer. After him, we know of only few attempts to analyze and interpret population data before the nineteenth century, in which new population sciences such as economics, evolutionary biology, sociology, demography, anthropology, statistics, and epidemiology surged. All these new sciences were built on the premise that there were issues that could only be studied using groups of people (populations) as the unit of analysis. Consider Thomas Malthus's (1766–1834) theory of societal implosions, Charles Darwin's (1809–1882) theory of natural selection, or Karl Marx's (1818–1883) theory of profit. They can only be explained in terms of populations.

2.3.3 Contagionism

Those who first used population studies were inclined to adhere to modern theories of disease causation, which singled out necessary and sufficient causes: if the cause was unique and the same for all, it was logical to compare groups of people who were exposed to it with those that were not. In contrast, the miasma theory was incompatible with population studies: if disease stemmed from a different set of causes in each person, there is no sense in grouping distinct individuals, just as one would not group apples and pears and call them the same fruit.

Thus, causal hypotheses based on the theory of contagionism, aka the germ theory of disease causation, could be tested using population studies. Even though no organism had yet been linked to a specific disease, contagionism had its minority of followers, one of whom was John Snow. Indeed, what Snow had in mind when he compared the frequency of cholera deaths among clients of the Southwark and Vauxhall, who pumped cholera-contaminated water, with those of the Lambeth companies, who pumped cleaner water, was simply that the postulated cholera germ present in the water was the necessary and sufficient cause. To avoid criticism from miasmatists, Snow was happy to report that both groups of clients were comparable in terms of exposure to air pollution.

The irony is that these cholera outbreaks in London would have remained localized if the sanitarians had not washed the streets of London into the Thames in the midst of the epidemic, creating optimal conditions for local outbreaks to be disseminated to the whole city (Johnson 2006).

Snow's investigation of the 1854 epidemic of cholera in South London does not correspond with any of the modern epidemiological designs. Snow's investigation consisted in categorizing the cholera deaths into clients of the Southwark and Vauxhall or of the Lambeth companies. He did not know the total number of clients supplied by each of the two companies in the districts in which he had enumerated and located the cholera deaths. He therefore used the total number of London houses supplied by the two companies in 1853 (40,046 and 26,107, respectively) as denominator for the company-specific numbers of deaths. Snow's work is nonetheless a large-scale use of both population thinking (i.e., a ratio of deaths over houses) and group comparison (i.e., clients of the Lambeth vs. Southwark and Vauxhall).

2.4 Preformal Epidemiology

Preformal epidemiology refers to the phase during which investigations combining population thinking and group comparisons were performed spontaneously, without guidance from an underlying theory. Epidemiology was not a *formal* scientific discipline yet.

I have argued previously that there could not be any form of epidemiology before the seventeenth century when population data regarding health-related issues became available (Morabia 2004, 2013). Soon after, however, group comparisons were conceived. William Petty (1623–1687), the English author of *Political Arithmetic,* and an early advocate of population data collection, imagined that he could challenge his colleagues in the Royal College of Physicians by asking questions such as whether they took as much medicine and remedies as "the like number of men of other societies," whether 1,000 patients of the best physicians would experience the same mortality rate as 1,000 patients of places where there dwell no physicians, whether 100 patients sick with acute diseases who used physicians would experience the same mortality rate as 100 patients who used no doctors and relied on chance only (Petty 1927, pp. 169–170). Each of these questions can only be answered by group comparison.

A group comparison involving purposefully collected data was conducted by James Lind (1716–1794) on the HMS Salisbury in 1747. Of the six pairs of scorbutic seamen receiving different treatments, those who received lemons and oranges were the only ones to be rapidly cured (Lind 1753). This small trial produced valid results where innumerous trial-and-error attempts to use different sorts of treatments, one patient at a time, had failed.

The other pioneering preformal epidemiological works have been abundantly described. This includes the work of the medical quantitativists in eighteenth century United Kingdom (Troehler 2000), Pierre Louis's (1787–1872) demonstration of the inefficacy of bloodletting as a treatment of pneumonia (Morabia 1996; Louis 1836), Ignaz Semmelweis's (1818–1865) 1848 intervention in Vienna in which he forced his interns to disinfect their hands before attending parturient women (Carter 1983), and John Snow's (1813–1858) work discussed above (Vinten-Johansen et al. 2003; Snow 1855, 1936).

A definition of epidemiol in contrast, the role of preformal epidemiologists in the investigation of outbreaks of infectious diseases has been understated (Morabia 1998). Epidemiologists were commonly called to investigate local outbreaks of diseases such as typhoid fever. In this activity, they could not rely on bacteriological evidence because bacteriology was still in its infancy in the late nineteenth century, and laboratories were not universally available. Using essentially survey and surveillance data, group comparisons, and detective skills, epidemiologists were able to show, for instance, that oysters could be the culprit of local outbreaks of typhoid fever (Morabia and Hardy 2005; Hardy 1999) and that cow's milk or condensed milk could cause fatal diarrhea in babies (Morabia et al. 2013).

Preformal epidemiological studies are characterized by the identification of a single causative factor. They come from people who were contagionists or believed that treatment had to be determined according to characteristics of the diseases, (e.g., scurvy, pneumonia, and cholera) and not by the unique constitution of individuals. These modern thinkers were few in number yet achieved a lot. There is not much knowledge about prevention or treatment that we credit pre-twentieth century science with but a great deal of it has been generated by preformal epidemiologists.

It is striking that Lind, Louis, Semmelweis, and Snow's conclusions were met with tremendous skepticism during their lifetime (Vandenbroucke et al. 1991). The main obstacle was the lack of agreement on whether or not groups were meaningful entities and, if they were, whether the groups in those studies were comparable or not. Critics saw a strong potential for what we now call confounding effects in the preformal epidemiological studies. Preformal epidemiologists were not methodologically and conceptually equipped to address these criticisms. A theory of group comparison was badly needed to support the design of stronger studies.

An important episode related to the history of the concept of interaction occurred during the nineteenth century. The German hygienist Max von Pettenkofer (1818–1901) attempted to reconcile miasmatism and contagionism into a theory of cholera origin, stating that the cholera germ had to be modified by its migration across porous soils in order to become a pathogenic miasma (von Pettenkofer 1869). Von Pettenkofer was trying to reconcile the old and new ways of thinking since interactions were at the core of holistic medicine and germs were a perfect case of monocausation. He failed under dramatic conditions mainly because he did not test his theory using group comparison (Morabia 2007a).

A definition of preformal epidemiology is provided in the description of the aims of the London Epidemiological Society was instituted in 1850: "the study of Epidemic and Endemic Diseases, with special reference to the investigation of (a) the various external or physical agencies and the different conditions of life which favour their development or influence their character; and (b) the sanitary and hygienic measures best fitted to check, mitigate, or prevent them. (...) While the chief object of the other medical societies of the metropolis is the investigation of the physiological, pathological, and therapeutic relations of diseases, that of the Epidemiological Society is specially the study of their etiological or causal relations, and the influences of locality, climate and season, diet and occupation, etc., on their rise, dissemination, and continuance. Diseases are looked at not so much in detail as in the aggregate; not in individual cases, but in groups and successions of cases; and not in one place only, but over wide and varied areas of observation. It is as a branch of Natural History rather than of technical medicine that the Society seeks to render the knowledge of epidemics and endemics more accurate and complete." (Anonymous 1876)

2.5 Early Epidemiology

Early epidemiology corresponds to the phase during which elements of an epidemiological theory were produced, epidemiology became an academic discipline, and epidemiological textbooks became available.

2.5.1 Population Thinking

2.5.1.1 Risk Versus Rate

In Snow and Louis's work, the occurrence of events in their studied population was expressed in a way that today looks naive. Snow used a ratio of cholera deaths over client houses. Louis used simple proportions. William Farr (1807–1883) who had a more mathematical mind than Snow and Louis, however, felt the need to formally distinguish risk and rates when trying to explain why cholera excited more terror than tuberculosis even though one had a 50% chance of surviving cholera vs. almost no chance of surviving tuberculosis (Farr 1838). The median survival time after diagnosis was 7 days for cholera and 2 years for tuberculosis. It was more frightful to have a 50% chance of death within the following few days than to be sure to die within the next 2 years. Risks were 46% over 7 days for cholera vs. 100% over 2 years for tuberculosis. What was needed was a measure that expressed the speed at which a disease killed. This was the rate. Farr expressed the risk by unit of time: 47% over 7 days became 47 per 100 per week, and 100% over 2 years became about 1.0 per 100 per week. The mortality rate from tuberculosis was about one fiftieth of that from cholera, reflecting the reasonable terror cholera excited in populations.

2.5.1.2 The Typhoid Vaccine Controversy

The concepts of risk and rate were further refined when Austin Bradford Hill (1897–1991) showed that it could be computed using person-times in the denominator. Hill's clarification occurred at the end of a long controversy between an army commission and a British team made of an epidemiologist, Major Greenwood (1880–1947), and a statistician, Udny Yule (1871–1951). Both groups diverged on the mode of analyzing the data from the antityphoid vaccine campaigns conducted in the British army (Susser 1977). Soldiers were inoculated at different points in time after the beginning of the campaign. The army commission wanted to define every soldier as inoculated, as long as they were inoculated by the end of the campaign. By doing so, on the one hand, they attributed some non-inoculated time free of disease to the vaccine, making it look better than it really was. Greenwood and Yule, on the other hand, wanted to define the inoculated as those solely inoculated at baseline, considering those inoculated later during the campaign as non-inoculated. In doing so, they made non-inoculation look better since inoculated-disease free time was analyzed as non-inoculated. Incidentally, both approaches showed a protective effect of the vaccine. But Hill extended the two ideas to a more accurate measure, showing that one could avoid the limitations of both approaches by computing person-times and allocating the follow-up period fairly into inoculated and non-inoculated time (Hill 1939).

2.5.2 Group Comparisons

Early epidemiologists invented the basic study designs currently used by epidemiologists.

2.5.2.1 Weinberg and Lane-Claypon Historical Cohort Study

The historical cohort study performed by Wilhelm Weinberg (1862–1937), in Germany, remains, by current standards, an impressive endeavor. Many people in public health had begun to worry by 1900 that the improved survival observed among tuberculosis people, probably due to a combination of better nutrition and isolation in sanatoria, would lead to a wide dissemination of the disease, which ultimately could jeopardize the "quality of the race." This mode of thinking was called eugenics in England and the USA, and racial hygiene in Germany. Weinberg performed his study to determine whether tuberculosis could threaten, in the long run, the average population health. Weinberg wanted to assess the fertility of the tuberculous and the survival of their offspring. He knew, however, that he could not address these questions in a cross-sectional survey since the only cases would be people who had survived after having contracted tuberculosis. These may be different from those who had died from tuberculosis. To enumerate all the cases occurring in the population, he realized that he had to follow people from birth. He therefore conceived an astute way of linking population and death registries of his state, Baden-Württemberg, and generated two cohorts: one comprising all the children (n = 18,052) of all the people who had died from tuberculosis between 1873 and 1889 and another cohort all the children (n = 7,574) of a sample of the people who had died from other causes during the same period. His study showed that those with tuberculosis had fewer children than those without tuberculosis and that their children died at a younger age (Weinberg 1913; Morabia and Guthold 2007). Weinberg concluded that tuberculosis was not a threat for the German race.

Weinberg's historical cohort study was published slightly after a much smaller one by Janet Lane-Claypon (1877–1967) in Berlin. Using the registry of a family planning clinic, she was able to show that breastfed infants grew up faster than infants fed with bottled milk (Winkelstein Jr. 2004).

2.5.2.2 Frost's Generational Cohort

Another ice-breaking historical cohort was conducted in the USA by Wade Hampton Frost (1880–1938) (Frost 1933). Frost's method of constructing the cohorts was novel. Starting with a survey of 556 persons living in 132 families from a small town in Tennessee, he identified, by interview, another 238 former members of the family who were dead or living. Altogether, they represented almost 10,000 person-years.

He then categorized the person-times according to the history of family contact with pulmonary tuberculosis. He found, as Weinberg before him, that the rate of tuberculosis infection was twice as high among those with a family contact as among those without such contact.

2.5.2.3 Goldberger and the Missed Nobel Prize

Another outstanding epidemiological work of the period was conducted by Joseph Goldberger (1874–1929) and Edgar Sydenstricker (1881–1936) in South Carolina (Goldberger et al. 1920). In an attempt to separate the effect of diet from that of income in the etiology of pellagra, they surveyed cotton mill workers and their families and followed them during the summer, monitoring the cases of pellagra. They found that lack of diversified diet was the culprit, and it was the consequence of insufficient income and a lack of access to affordable fresh food forcing the pellagrous-prone families to subsist predominantly on corn. Goldberger should have been awarded a Nobel Prize in 1929 as part of a series of awards related to the discovery of vitamins, but he died that same year.

2.5.2.4 Case-Control Studies

Early epidemiologists conducted case-control studies. The one reported in 1926 by Janet Lane-Claypon stands out as a remarkable and innovative project for its design and magnitude (Lane-Claypon 1926; Morabia 2010; Press and Pharoah 2010). She compared 500 hospitalized breast cancer patients with 500 control subjects, hospitalized for non-cancerous illnesses. All were interviewed the same way. The study showed that breast cancer patients had fewer children than the control subjects, even though, from a modern perspective, both groups were hyperfertile, with an average of 3.5 children among the breast cancer patients and 5.5 children among the control subjects. In this study, Lane-Claypon mentioned that cases may recall their past exposure differently than the controls, a potential threat to the validity of case-control studies now known as recall bias (Lane-Claypon 1926).

Most case-control studies subsequent to the 1926 study but prior to 1945 demonstrated a growing understanding of the strengths and pitfalls of case-control studies. In 1928, Lombard and Doehring matched each of their cancer records with the record of an "individual without cancer, of the same sex and approximately the same age" (Lombard and Doering 1928). The 1931 study by Wainwright (1931) reanalyzed by Press and Pharoah (2010) replicated Lane-Claypon's 1926 study. In 1933, Stocks and Karn filled two full pages of the *Annals of Eugenics*, describing their matched case and control selection, stating in particular that "the control series is not a random sample, but is made to conform to the case series, which is a random sample of a cancerous population" (Stocks and Karn 1933). At least three case-control studies were published before 1945 about the determinants of ischemic heart disease and hypertension (Johnson 1929; Glendy et al. 1937; English et al. 1940). They all provided detail about the comparability of the controls to the cases.

The 1939 (Mueller 1939a, b) and 1943 (Schairer and Schoeniger 1943, 2001) German studies on smoking and lung cancer had designs below the standard of their time (Morabia 2012, 2013b). Mueller described 86 cases of lung cancer at length, some deceased and some alive, only to say that they were compared with "the same number of healthy men of the same age" (Mueller 1939b). Schairer and Schoeniger compared 93 deceased cases of lung cancer, 226 deceased cases of other cancers, and 270 "controls," all men (Schairer and Schoeniger 1943, 2001). They mentioned only that the controls were men from Jena, "aged 53 and 54 years," and that 270 out of the 700 invited returned satisfactorily filled questionnaires. These studies marked a regression in quality and rigor relative to the pre-Nazi German epidemiology of Weinberg (Morabia and Guthold 2007; Weinberg 1913).

2.5.3 Concepts

2.5.3.1 Confounded Effects as an Artifact

Epidemiologists of the first half of the twentieth century searched for formal ways to address the criticism of non-comparability (Morabia 2007b, 2011). They implemented new techniques such as random allocation of treatment (Therapeutic Trial Committee of the Medical Research Council 1934) and restriction of the study sample (Goldberger et al. 1920). They improved the epidemiological study designs with historical cohort studies (Weinberg 1913; Winkelstein Jr. 2004) and case-control studies (Lane-Claypon 1926; Lombard and Doering 1928; Johnson 1929; Stocks and Karn 1933; Glendy et al. 1937; English et al. 1940). Analytical techniques were used to "standardize" compared risks and rates (Goldberger et al. 1920; Weinberg 1913) and used exposure propensity score equation (Lane-Claypon 1926). All of these efforts had the same objective: to design studies and analyze data in ways that purposefully optimized comparability of exposure to potential alternate causes of the studied outcome.

Surprisingly, the first definition of the concept we refer to today as "confounding" did not follow from this line of efforts of balancing the effects of multiple independent causes across groups.

2.5.3.2 The Pooling Bias

In 1903, G. Udny Yule, the Cambridge statistician, gave an early description of a form of bias, or "fallacy", which could lead to spurious associations (Yule 1903). He showed that there were situations in which pooling strata could provoke the appearance of associations which did not correspond to the effect observed within each stratum. The same fallacy was independently described 36 years later by Major Greenwood, the first Professor of Epidemiology at the London School of Hygiene and Tropical Medicine (Greenwood 1935), (pp. 84–85), and 40 years later by Austin Bradford Hill (Hill 1939), (pp. 125–127). Yule, Greenwood, and Hill warned analysts against the dangers of pooling data. They understood that group non-comparability after pooling strata was the underlying problem but did not describe the fallacy as a major source of bias in population studies and did not suggest computing weighted averages of the stratum-specific effects to bypass the fallacy.

2.5.3.3 Randomization

During this period randomized controlled trials were conducted. Under the conceptual leadership of Hill, the Medical Research Council (MRC) of the UK conducted a trial to assess the efficacy of serotherapy in the treatment of pneumonia (Therapeutic Trial Committee of the Medical Research Council 1934). Patients were allocated to either serum rich in antipneumonia antibodies, or rest. They were randomized by alternate allocation, with every other patient receiving serotherapy. However, the allocation procedures were too transparent for the clinicians who seemed to have had a tendency of preferentially giving serotherapy to younger patients with better prognoses. This flaw led later experimenters to conceal the randomization procedures (Chalmers 2002), as in the randomized control trial conducted by Joseph Asbury Bell in 1941 (Chalmers 2010). This student of Frost tested the efficacy of a pertussis vaccine.

2.5.4 Early Epidemiology

The early epidemiology phase lasted up until 1945. Early epidemiologists comprised Weinberg, Lane-Claypon, Frost, Greenwood, Hill, Goldberger, Sydenstricker, and Bell. They conducted the first historical cohort studies, case-control studies, and randomized controlled trials. They identified ways to improve comparability in observational studies.

Epidemiology became academic. Frost was appointed first American Professor of Epidemiology, after having been "Resident Lecturer" since 1919. Greenwood became the first Professor of Epidemiology in the UK in 1928. Epidemiology was defined, but the definition was evolving rapidly. Frost's first definition reflected his own training as a public health officer: epidemiology was the study of the natural history and the causes of infectious diseases. In 1924, Frost expanded the definition to include tuberculosis and "mass phenomena of such non-infectious diseases such as scurvy," but he excluded the "so-called constitutional diseases such as arteriosclerosis or nephritis" (Comstock 2004). In 1935, Greenwood further expanded the definition to include "any disease" which manifested itself as a "mass phenomenon" (Greenwood 1935).

In 1935, Greenwood published a textbook of epidemiology, although there is some controversy surrounding its qualification as a "first" textbook (Bracken 2003). In 1941, Frost published a large article, entitled Epidemiology, which reads like the outline of a more extensive textbook (Frost 1941).

2.6 Classic Epidemiology

2.6.1 Tobacco and Lung Cancer Before Epidemiology

The uptake of cigarette smoking in Western societies throughout the first half of the twentieth century remains a puzzling public health phenomena. In less than 50 years, a majority of men became chronic smokers. At the time, most of the public only saw cigarettes as a benign product, at worst (i.e., what can a little bit of smoke do to your lungs?), and, at best, a product that kept them healthy, as Marlene Dietrich (1901–1992) is famously known to have claimed. Still, some researchers suspected that it could have deleterious effects on human health. By the end of the 1920s, the Argentinean Angel Roffo (1882–1947) had hypothesized, supported by laboratory and animal experiments, that tobacco could cause cancers (Proctor 2006). In 1938, the statistician Raymond Pearl (1879–1940) reported a 10-year shorter life expectancy of heavy smokers vs. light or non-smokers on the basis of a life table analysis of 6,813 white males (Pearl 1938). Thoracic surgeons in particular were surprised by the rise in lung cancer patients, a very rare occurrence until 1900. In 1939, Alton Ochsner (1896–1981) and Michael DeBakey (1908–2008), in a report of 79 cases of lung cancer, mentioned “smoking” as “undoubtedly a source of chronic irritation of the bronchial mucosa” and therefore a possible cause of lung cancer (Ochsner and DeBakey 1939). They provided, however, no data on the smoking habits of their cases and had no control group. They actually attenuated this statement in their second series of 129 cases published in 1947. Having shown that 75% of the subjects smoked, they enigmatically concluded that smoking habits “seemed of no particular significance in this particular series” (Ochsner et al. 1947). Once again, they had no control series.

In the background, incidences of lung cancer were soaring. Medical societies were debating whether it was real or “only apparent” because of improved diagnosis due to the invention of x-rays (Medical Research Council 1952). On the basis of the San Francisco Cancer Survey, Frederick Hoffman, in 1931, had already concluded that tobacco caused cancer of the lung (Hoffman 1931), but his results, further confirmed by the two German studies of 1939 and 1943, had little impact on the debates. It is unclear how much researchers in the UK and the USA knew about these studies before the war. However, given the favorable political climate surrounding cigarettes and in the light of the obstacles encountered in the publication process of much better studies immediately after the war, it is unlikely that these studies could have weighed in the balance of the pros and cons about the carcinogenic effect of tobacco.

New evidence was provided by case-control studies published in 1950 in the USA (Wynder and Graham 1950; Levin et al. 1950) and in the UK (Doll and Hill 1950). Richard Doll (1912–2005) and Austin Bradford Hill’s studies represented major methodological developments compared to the pre-World War II epidemiology.

2.6.2 Group Comparisons

2.6.2.1 Case-Control Studies

Doll and Hill compared 647 male cases of lung cancer recruited from 20 hospitals in London with 647 male controls, who were patients admitted to these same hospitals for non-cancer conditions (Doll and Hill 1950). They were all interviewed about their past smoking habits, including duration, date of starting and quitting, amount smoked, and type of tobacco.

The prevalence of smoking was 99.7% among cases and 95.8% among controls. Even when one considered the amount smoked, differences were not startling. But because the study was population-based, Doll and Hill were able to divide, within age categories, the number of cases in the study, by the size of the Greater London population represented by the controls. For example, if 10% of the controls were light smokers, they assumed that 10% of the male residents of Greater London were light smokers at the time of the study. Dividing the number of light smoker cases by whichever number, 10% of the male population provided an approximation of the mortality rate of light smokers. The mortality rate of a category of smokers further divided by the mortality rate of non-smokers would represent a "ratio," which Doll and Hill thought approximated the corresponding "relative risk" (Doll and Hill 1950, p. 491). For example, in the age group 45–54, smokers of 15–21 cigarettes per day had 20 times the risk of lung cancer as non-smokers; those smoking 50 cigarettes per day had a relative risk of 55.6.

Thus, Doll and Hill performed a case-control study as a way to estimate risks and relative risks. They did not use odds ratios, which became the preferential way of approximating risk ratios in case-control studies rapidly after.

2.6.2.2 Cohort Studies

One year after their case-control study was published, Doll and Hill began the British Doctors Study. They had decided to perform a "prospective study" because they thought that "if there was any undetected flaw" in case-control (they used the term "retrospective") studies, "it would be exposed only by some entirely new approach" (Doll and Hill 1954). They sent a very short questionnaire to the 59,600 doctors of the United Kingdom about their smoking habits. They got 40,564 valid answers from doctors (34,445 men and 6,192 women) and started tracking their dates and causes of death beginning on November 1, 1951. Their preliminary report in 1954 (Doll and Hill 1954) indicated that the "standardized rates from cancer of the lung for 1,000 men aged 45–75 years in relation to the most recent amount of tobacco smoked" were very similar when "estimated from (a) the "backward" inquiry into the history of patients with cancer of the lung and other diseases in London and (b) from the present "forward" inquiry into the mortality of doctors" (Doll and Hill 1954). In modern terminology, Doll and Hill's cohort and case-control studies had provided very similar rate ratios for the association between cigarette smoking and lung cancer.

The British Doctors Study was followed by other larger cohort studies in the United States. The American Cancer Society (ACS)'s "Hammond and Horn Study" was launched in January 1952. Its leaders, E. Cuyler Hammond (1912–1986) and Daniel Horn (1916–1992), had conceived a convenient way to rapidly involve 200,000 Americans (Hammond and Horn 1958a, b). They recruited ACS members and their relatives and friends. The study confirmed that smoking had a stronger association with lung cancer than with coronary artery disease yet also demonstrated that, in terms of excess deaths, many more smokers died of coronary artery disease than of lung cancer (Hammond and Horn 1958b). In 1953, Harold Dorn (1906–1963) launched an even larger cohort study than that of the ACS. The US Veteran's cohort study enrolled 300,000 male holders of active life insurance policies from the Veteran Administration (Dorn 1959).

2.6.2.3 Cornfield's Rare Disease Assumption

Jerome Cornfield (1912–1979) has made key contributions to epidemiological methods and concepts. His biography has yet to be written. In 1951, he demonstrated that the risk ratio could be approximated by the ratio of the odds of exposure in cases over the odds of exposure in controls derived from a case-control study (Cornfield 1951). In doing so, Cornfield provided the solution to one of the main obstacles to recognizing case-control studies as valid for assessing causal associations. Before Doll and Hill's case-control studies were only viewed as able to determine whether cases had experienced different exposures than controls (Lane-Claypon 1926; Lombard and Doering 1928). This was an indirect answer to the causal question: does exposure impact the risk of disease? Doll and Hill had assumed that the controls were representative of all residents in the Great London to compute "ratios." Cornfield perceived that, using Bayes theorem, it was possible to transform a conditional probability of exposure derived from case-control studies, into a conditional probability of disease, which was the causal question. The Bayesian derivation of the probabilities themselves required an estimate of the rate of the outcome, say a disease, in the studied population, information that case-control studies did not provide. Nonetheless, using Bayes theorem, the ratio of conditional odds of exposure in cases and controls, approximated a ratio of conditional probabilities of disease in exposed vs. non-exposed, that is, a risk ratio, given that the rate of the disease in the population was low. This key condition for the odds ratio (a term not used by Cornfield) to approximate the risk ratio is known today as the "rare disease assumption." Using data from a case-control study of smoking and lung cancer by Robert Schrek (dates unknown) and coauthors (Schrek et al. 1950), Cornfield demonstrated that the rate ratio of lung cancer could be approximated by the ratio of the odds of exposure since the annual prevalence of lung cancer in the population was only 15.5 per 100,000. Cornfield also added the caveat that, for this approximation to be valid, cases and controls had to be representative of all cases and of all people free of disease in the target population.

2.6.3 Concepts

2.6.3.1 Confounding

The classic concept of confounding is paradoxically unrelated to the work of early epidemiologists but has its roots in a controversy about the statistical assessment of interaction (Vandenbroucke 2004). The British statistician Edward H. Simpson explored whether the interaction term could be left out of a mathematical model if the stratum-specific odds ratios were similar (Simpson 1951). His conclusion, known today as Simpson's paradox, was that this was possible only if there was no third variable related to outcome among the non-exposed, and to exposure among the non-outcomes. This paradox became the basis for the classic definition of a confounding variable, or confounder, probably through the work of sociologist Leslie Kish (1910–2000), who linked the paradox with the qualifier "confounding" (Kish 1959; Vandenbroucke 2004).

In 1959, Nathan Mantel (1919–2002) and William Haenszel (1910–1998) published their influential estimates of the adjusted odds ratio (Mantel and Haenszel 1959). They did not use the words odds ratio, or any form of the term "confound," but the paper was about computing an average estimate of the odds ratio across strata of a confounding variable.

The qualifier confounding, as well as variations of Simpson's paradox, appears in the epidemiological literature from 1970 on (MacMahon and Pugh 1970; Susser 1973). The hypothetical examples provided in textbooks illustrated situations in which the odds ratio and the rate ratio were homogeneous across the strata of a confounder but different from when strata were pooled to compute a single crude odds ratio. These examples were identical to those given by Yule, Greenwood, and Hill in the previous phase of epidemiology history, but no explicit reference to these early epidemiologists was made (Morabia 2011).

2.6.3.2 Interaction

Hammond has made many insightful contributions to epidemiological methods and concepts. Along with Dorn, he had conceived the first cohort study sponsored by the American Cancer Society. In his book chapter on *Causes and Effect* (Hammond 1955), Hammond had assembled a list of characteristics that a causal association could satisfy which is very similar to the list published 10 years later by Hill (1965).

But Hammond, teaming up with Irving Selikoff (1915–1992), also paved the way for a modern and quantitative assessment of interactions (Hammond et al. 1979). Selikoff had pioneered the research on the health effects of asbestos for which he had assembled a large cohort of insulators and other asbestos workers, all of whom were exposed to asbestos. To serve as unexposed comparison, Hammond selected within the ACS cohort study blue-collar workers exposed to dust and fumes. Comparing asbestos workers with the ACS cohort, they were able to show that smoking and asbestos both caused lung cancer in terms of mortality rate differences and rate ratios but that their joint rate difference was much larger than the sum of their individual rate differences. They concluded that the presence of a "synergistic effect" meant that "a young person who is so strongly addicted to cigarette smoking that he cannot break the habit or is unwilling to do so would be particularly well advised not to enter a trade involving exposure to asbestos dust."

2.6.3.3 Bias

The origin of the classic theory of selection bias is justly attributed to Joseph Berkson (1899–1982), a statistician at the Mayo Clinic. In the 1940s, Berkson had worked out a mathematical example, which aimed to illustrate how comparative studies based on autopsy records could generate spurious associations strictly on the basis of the selection probabilities of the diseases. He had been prompted to this by a 1929 publication of Raymond Pearl (1879–1940), a Johns Hopkins statistician, which suggested that tuberculosis protected against cancer since it was less commonly found at autopsy among people who had died from cancer than among those who had died from other causes (Pearl 1929b). Berkson showed mathematically that the bias was unavoidable under Pearl's study design (Berkson 1946). His point was that if the cases had disease A (e.g., cancer), controls had disease B (e.g., coronary heart disease), and the "exposure" was disease C (e.g., tuberculosis), the simple mix of conditional probabilities of admission for exposed and non-exposed cases and controls would generate a biased effect. Later, Berkson, a skeptic of the tobacco-lung cancer association, attempted to generalize the domain of application of his selection bias from autopsy studies to all types of case-control and cohort studies (Berkson 1955). Berkson's theoretical demonstration became the model for selection bias in epidemiology training, even though it has proved extremely difficult to illustrate the bias empirically (Roberts et al. 1978).

2.6.3.4 Misclassification Bias

The foundation for a classic definition of misclassification bias was also established in the 1950s. In 1954, Ernst Wynder (1922–1999) et al. had reported that circumcision of the husband protected the wife from cervical cancer (Wynder et al. 1954). A few years later, in 1958, Lilienfeld and Graham challenged the validity of the information wives could provide about their husband prepuce coverage (Lilienfeld and Graham 1958). They supported their argument by showing that even men were inaccurate about their own circumcision status. In a comparison of self-reported circumcision statuses of 192 men with the clinical evaluation of a doctor, Lilienfeld and Graham found that 56% of the self-reports were a false negative and 18% were a false positive. A debate followed on the impact of misclassification of exposure on the estimation of effects (Diamond and Lilienfeld 1962a, b; Keys and Kihlberg 1963). The terms of the debate were clarified by Newell (1962). Later, Copeland et al. (1977) summarized the impact of non-differential and differential misclassification on measures of effects and provided ways to correct for the bias when the misclassification magnitude (e.g., sensitivity and specificity) was known.

2.6.3.5 Classification of Biases

The 1950s, 1960s, and 1970s saw a proliferation of biases that could occur at any level of an epidemiological study. David Sackett took a stab at a bias typology. Having identified 35 sources of biases, he classified them according to whether they occurred in the design, sample selection, conduct, measurements, analysis, or interpretation of a study (Sackett 1979). Classic epidemiology textbooks do not go beyond mentioning different sources of bias.

2.6.4 Classic Epidemiology

The classic definition, popularized by Last's *Dictionary of Epidemiology*, states that epidemiology is "the study of the distribution and the determinants of health-related states of events in specified populations, and the application of this study to control health problems" (Last 2001). The definition weaves descriptive epidemiology ("distribution"), analytical epidemiology ("determinants"), and public health ("application") together.

Textbooks flourished, written by leading researchers (Morris 1957; MacMahon et al. 1960; MacMahon and Pugh 1970; Susser 1973; Lilienfeld 1976).

The symbolic achievement of classic epidemiology was the publication of the 1964 US Surgeon General report, *Smoking and Health* (The Surgeon General's Advisory Committee 1964). The report had been prepared by a committee comprising only people that had been unanimously proposed or tolerated by the tobacco industry. It relied essentially on epidemiological evidence and concluded that there was a causal link between cigarette smoking and lung cancer in men.

2.7 Modern Epidemiology

2.7.1 Population Thinking

2.7.1.1 Estimability and Estimation

Miettinen's "Estimability and Estimation in Case-Referent Studies" (Miettinen 1976) most likely marks the transition from classic to modern epidemiology for two reasons: (1) it explicated the link between risk and rate, respectively, baptized cumulative incidence and incidence density; (2) it unified the two major epidemiological study designs in generalizing the link between cohort and case-control studies. The core of Miettinen's model is depicted in the now historical (one could even say "cult") Fig. 2.1 of Miettinen 1976 (see also Fig. 2.1 in this chapter).

2.7.1.2 Risk and Rate

The way Miettinen approaches risk and rate strikes me to be analogous to the way William Farr described them about 150 years earlier. Farr said that prognosis could be predicted "with certainty," as a rate, in populations, and predicted with risks as "probabilities," in individual patients (Farr 1838). Miettinen wrote, "for populations it is desired to learn about rates," and "for individuals the concern is with risks (for various time periods)" (Miettinen 1976). Miettinen extended Farr's description by deriving the mathematical relationship between rates and risks. Dividing the number of incident cases by the number of person-times for a very short period of time provided an instantaneous risk, called "incidence density." Summing or integrating the instantaneous risks over a longer period of time provided a cumulated risk, called "cumulative incidence.'' Four years later, Morgenstern, Kleinbaum, and Kupper translated this theory in simpler terms and applied examples (Morgenstern et al. 1980).

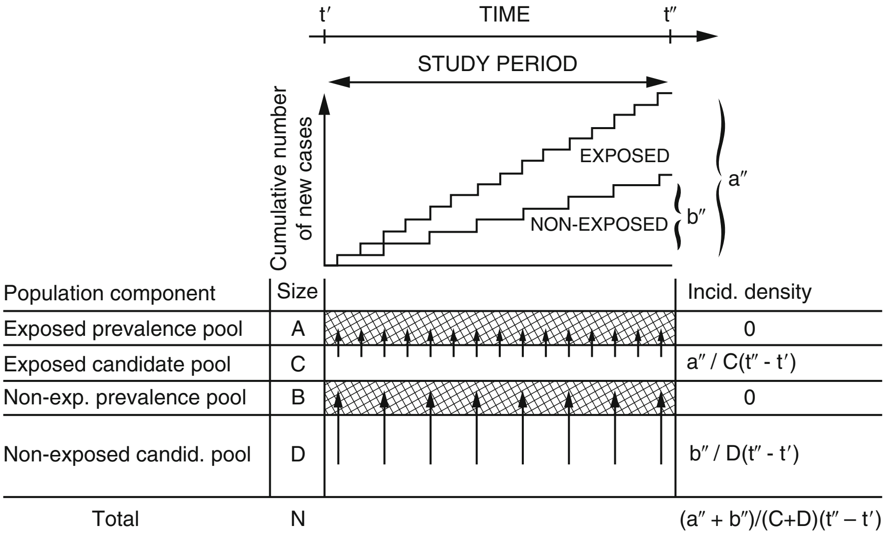


Fig. 2.1 Static population (e.g., a particular age group) over the time span of a case-referent study based on *incident* cases. The sizes of the different component populations remain static, but there is turnover of membership in each compartment. The *arrows* indicate occurrences of new cases, i.e., transitions from the candidate pools to the prevalence pools. Note that the incidence densities are zero in each of the prevalence pools and that the incident cases are referred to the follow-up experiences in the candidate pools only. Also note that the incidence density ratio is (a″/b″)/(C/D), with a″/b″ and C/D estimable from the (incident) cases and referents, respectively, regardless of the levels of incidence or prevalence (Source: Miettinen 1976)

2.7.2 Group Comparison

2.7.2.1 Rare Disease Assumption Revisited

In "Estimability and Estimation in Case-Referent Studies," Miettinen described the assumption formulated by Cornfield in 1951 according to which the odds ratio could approximate the risk ratio *only if* the disease was rare and the target population as "overly superficial and restrictive" (Miettinen 1976). It depended, Miettinen argued, on the way controls were selected from hypothetical underlying cohorts of exposed and non-exposed at-risk subjects. The "rare disease assumption" was appropriate in situations in which the controls were sampled from subjects who had not developed the outcome by the end of the risk period, that is, after all the cases had been identified. "No rare disease assumption was however required" for the situation in which the controls were sampled from the whole cohorts at baseline, or from the people still at risk of the outcome by the time the case was recruited into the study. Actually, the odds ratio could potentially be estimated to be exactly the same as the risk ratio when controls were sampled from the base and exactly the same as the rate ratio when controls were sampled concurrently to the cases, however common the disease was. These elements of Miettinen's paper were further explicated by Greenland and Thomas (1982) and by Greenland et al. (1986).

2.7.3 Concepts

Modern epidemiologists refined the concepts of confounding, interaction, and bias.

2.7.3.1 Confounding

The notion of a confounder as an additional variable brought epidemiologists to specify, in the late 1970s and early 1980s, that, in addition to the two properties already formulated by Simpson (being associated with exposure in the population, and associated with the outcome among the unexposed), the third variable should not mediate the relation of exposure to outcome (Greenland and Neutra 1980; Rothman 1977; Miettinen and Cook 1981). Otherwise, the approach to the concept had not changed substantially between Yule's expression of confounding as, "a fallacy caused by the mixing of records [i.e., strata]" (Yule 1903), and Rothman's "On the simplest level, confounding may be considered as a mixing of effects" (Rothman 1986, p. 89).

The classic definition of confounding had weaknesses. It was derived from the relation of additional variables to exposure and outcome and not from the characteristics of the studied association, such as non-comparability. A variable could meet the classic definition and not be a confounder (Greenland and Robins 1986). Matching for a confounder had different implications in cohort and case-control studies (Miettinen and Cook 1981; Kupper et al. 1981). Screening for confounding by comparing the stratum-specific and the pooled effects could lead to different conclusions based on whether one used risk ratios, risk differences, or odds ratios (Miettinen and Cook 1981).

Of note, one of the rare empirical examples of strong confounded associations was published in Lancet, in 1982. The paper reported a 12-fold higher risk of Kaposi's sarcoma in homosexual men who had used aphrodisiac poppers 500 times or more in their life compared to those who used them less than 500 times (Marmor et al. 1982; Morabia 1995). It was first believed that the confounder was HIV, but the HIV-Kaposi's sarcoma association itself was confounded by a herpes virus associated with multiple partner intercourses connected to the frequent usage of poppers and exposure to HIV (Moore and Chang 1995).

Departure from the fallacy-based definition of confounded effects was initiated in a 1986 paper by Greenland and Robins, which made the potential outcome approach conceived by Splawa-Neyman (1990) and developed by Rubin (1974) widely accessible to epidemiologists (Greenland and Robins 1986). Their potential outcome model was confined to deterministic risks (i.e., risks that can equal only 0 or 1), and the explication was based on the four "causal" types described by Copas (1973): "doomed," "exposure causative," "exposure preventive," and "immune." Each causal type was characterized by two different outcomes (whether exposed or non-exposed) and each led to a different causal explanation. There were therefore two causal explanations for each observed effect *in an individual*. In group comparisons, however, the proportion of doomed and immune could be balanced and the causal effect partially or completely identified. The observed effect was however confounded when the doomed and/or immune causal types were not balanced across the compared groups (Greenland and Robins 1986). This theory of confounding derived from potential outcome contrasts was generalized from randomized to observational studies (Rubin 1990; Hernan and Robins 2006). It was also compatible with a fallacy-based definition since non-exchangeable groups differ on variables that meet the three criteria of confounding variables.

The foundations, contributions, and segues of the Greenland and Robins 1986 paper were recently reviewed by its authors (Greenland and Robins 2009). From a broader historical perspective, "identifiability, exchangeability, and confounding" marks the transition from a definition of confounded effects based on effect mixes, to a theory of group comparability.

In later developments, it was shown that comparability could be achieved by standardization, stratification, inverse probability weighting, and other techniques beyond randomization (Hernan and Robins 2006) and that connections could be drawn between potential outcome models, directed acyclic graphs and structural equations (Hernan et al. 2004; Greenland et al. 1999; Pearl 1995, 2000; Greenland and Brumback 2002). These new ways of analyzing data produced results of practical value (Cole et al. 2003; Hernan et al. 2000).

2.7.3.2 Interaction

Much thought has been invested in the concept of interaction by modern epidemiologists. A Medline search of the American Journal of Epidemiology and the International Journal of Epidemiology (started in 1972) showed that the term "interaction" was practically never used other than to describe infectious processes before Rothman's landmark 1974 paper (Rothman 1974). Since then, textbooks have provided a theory of how to assess, quantify, and test for interactions on multiplicative or additive scales (Kleinbaum et al. 1982; Rothman and Greenland 1998; Szklo and Nieto 2000; Templeton 2000; Susser et al. 2006).

Interaction is what glued components together in Rothman's causal pie (Rothman 1976) since a cause is not complete as long as one of its components is missing. Additional developments of the theory of interaction included the case-only studies (Yang et al. 1997) and the role of parallelism in assessing synergy (Darroch 1997). There are ongoing attempts to integrate the concept of interaction to causal models as was done for confounded effects (Vanderweele and Robins 2007).

2.7.3.3 Bias

The modern theory of bias went beyond Sackett's typology. Modern epidemiologists found that all biases fell under two broad categories, selection biases and misclassification biases. Most textbooks followed thereafter, in the dichotomization of bias sources (Miettinen 1985; Kleinbaum et al. 1982; Rothman 1986).

In 1977, Copeland et al. summarized the state of knowledge with respect to misclassification bias for associations involving two binary variables. The rule was that the bias would tend to underestimate the true effect if the degree of misclassification was similar in the compared groups, that is, if it was non-differential, but that it could either underestimate or overestimate the true effect if the information from one group was more misclassified than that from another.

Selection bias was thought for a while to be similar to confounded effects except that the confounder was measured (Rothman 1986). It was however later shown that confounding was attributable to exposure and outcome having a common "cause," while selection bias was due to exposure and outcome having a common "effect" (Hernan et al. 2004).

2.7.4 Modern Epidemiology

The definition of epidemiology provided in the most influential textbook of the phase states that "the ultimate goal of most epidemiological research is the elaboration of causes that can explain patterns of disease occurrence" (Rothman and Greenland 1998, p. 29).

Many textbooks have translated the modern epidemiological credo in their own ways (Kleinbaum et al. 1982; Rothman 1986; Rothman and Greenland 1998; Rothman et al. 2008; Szklo and Nieto 2000). Characteristically, one third of the textbook entitled *Modern Epidemiology* is dedicated to the application of the methods to specific domains of research such as nutritional, environmental, genetic, and clinical epidemiology (Rothman et al. 2008). This reflects a process of differentiation of field of expertise among epidemiologists associated with a flurry of specialized epidemiological textbooks.

In my view, the main contribution of modern epidemiology has been the assembly, homogenization, and unification of the diverse elements of theory existing about measures of disease occurrence, study design, confounding, interaction, and bias.

Building on the classic epidemiology tradition, modern epidemiology has continued to focus on isolating single causes of often complex, mostly non-communicable diseases, with apparently declining return. Modern epidemiology has been criticized for being "prisoner to the proximate" (McMichael 1999) focusing on a single level of organization (Susser and Susser 1996) favoring point vs. life course exposures (Kuh and Shlomo 2004) and neglecting complexity in favor of parsimony (Schwartz and Susser 2006).

2.8 Causal Inference

Historically, the combination of population thinking and group comparison has allowed epidemiologists to solve causal questions that were beyond the reach of holism and clinical management of single patients. It is therefore irrelevant to look for epidemiological modes of causal inferences prior to the emergence of population thinking in the seventeenth century. The lack of valid inferences related to health determinants, prevention, or treatment is obvious in the Hippocratic treatises as well as throughout the ancient medical literature (Morabia 2004).

The Scottish philosopher David Hume (1711–1776) in the eighteenth century. Hume proposed a set of "rules by which to judge of causes and effects" in situations in which the connection was not obvious, and in particular, when the cause was not deterministically followed by the effect (Hume 1978). These rules were logical criteria which one would expect to characterize a causal relation. For instance, the cause had to precede the effect; the effect intensity had to vary with the intensity of its cause; effects had specific causes and causes had specific effects; and analogous effects had to have analogous causes (Morabia 1991). But Hume's rules were not meant to be applied to evidence generated from group comparisons. They were therefore not directly relevant for epidemiology.

In a 1955 chapter of a book about smoking and health, E Cuyler Hammond listed a series of criteria to rule out the role of a "third factor" when attempting to relate a cause to its probabilistic effect (Hammond 1955).

All trained epidemiologists are familiar with Austin Bradford Hill's synthesis of the epidemiological approach to causality (Hill 1965). It had been preceded by a rich and international discussion among epidemiologists (Blackburn and Labarthe 2012). In contrast to Hume's rules, Hill's list of "viewpoints" applies specifically to evidence generated from group comparisons. It specified that, except for the cause preceding the effect, none of the criteria was indispensable. Hill also stressed that it was meaningless to combine criteria into scores because an association was not more likely to be causal if it met more criteria. Hill also explicitly mentioned that statistical tests were irrelevant for causal inference. It is of note that Hill did not use the term "criteria" at all but referred to "aspects," "features," "characteristics" of associations, or "viewpoints" from which we should study associations "before we cry causation."

Hill's synthesis has remained the major reference for causal inference in epidemiology. His paper was *not* replaced by later attempts to enrich it (Susser 1977) or supplanted by attempts to restrict causal inferences to observations obtained deductively rather than inductively (Buck 1975; Rothman 1988).

Causal inference is about synthesizing all we know at a given point in time, beyond the epidemiological results. Concluding that the human papilloma virus caused cervical cancer involves more than the epidemiological association of the virus and the disease (Muñoz et al. 1992). It incorporates the knowledge about the virus and about carcinogenic process derived from laboratory experiments. Causal inference goes beyond the strict context of epidemiological methods and concepts.

2.9 Epistemology of Epidemiology

To write this history of epidemiological methods and concepts, I used a mode of framing the evolution of scientific discipline referred to as "genetic epistemology" by Jean Piaget (1896–1980) (Piaget 1970; Morabia 2004). The idea is that scientific disciplines are in continual construction, formalization, and organization. Their methods and concepts are commonsensical when the discipline first appears but become increasingly theoretical and abstract as the discipline acquires experience and addresses questions of increasingly complex nature (Piaget 1970).

This work relies therefore on the premises that the methods and concepts that eventually formed epidemiology (a) had a history, which started with commonsensical comparisons and observations, (b) evolved as more abstract and formal with more rigorous methods and overarching concepts, and (c) are still evolving today. To trace this history, I have used the four phases (preformal, early, classic, and modern) previously identified in the history of epidemiological methods and concepts (Morabia 2004). I summarize here how each method and concept evolved across the four phases.

Population thinking. Risk and rate were used intuitively from Graunt to Snow to describe cumulative or instantaneous occurrences of health events. Farr distinguished them theoretically, and Miettinen, 150 years later, showed how risks could be derived from rates. The distinction and relationships between prevalence and incidence have been progressively worked across early, classic, and modern epidemiology.

Group comparison, before the twentieth century, lacked directionality and did not provide for group comparability. From 1900 on, study designs were refined, and techniques for reaching comparability were invented and used. In the 1950s, it was understood that cohort and case-control studies were mathematically linked. Later, it was shown that both designs could be conceptualized as different ways of sampling the same underlying cohorts.

Confounding was first expressed as a commonsensical issue of group non-comparability, later as an uncontrolled fallacy, then as a controllable fallacy resulting in confounded effects, and, more recently, as an issue of group non-comparability of potential outcome types. This latest development tied together the seemingly disconnected phases of the history of confounding as it established that group non-comparability was the essence of confounding, and that statistical fallacy was one of its consequences.

Interaction belonged initially to an early form of complex thinking, in which multiple causes had to be combined for an effect to occur. The classic concept of interaction has restricted the usage of the term to situations in which the combined effects of multiple causes are not the simple combination of their individual effects. Further theoretical developments have been hindered by the complexity of the subject and the difficulty of proving the empirical importance of interactions.

Bias was viewed as a collection of errors occurring at different levels of the epidemiological investigation until a consensus emerged that most biases could stem from selection or misclassification.

2.10 Conclusions

This review of the historical contribution of epidemiology to health knowledge indicates that epidemiology is characterized by the combination of population thinking and group comparisons aimed at discovering the determinants of human health, assessing the efficacy and side effects of treatments, and assessing the efficacy of screening. The set of methods (study designs) and concepts (measures of disease occurrence, confounding, interaction, and bias) have evolved since the seventeenth century and are therefore relatively novel (Morabia 2013a)

In this evolution, we can identify four phases characterized by qualitative leaps in formalization and abstraction of methods and concepts. After a preformal phase, in which epidemiology was discovered intuitively by scientists, and mostly physicians, epidemiology has gone through an early, a classic, and a modern phase.

The next phase in the evolution of epidemiological methods and concepts does not belong in this review, but it can be stated with confidence that the discipline is evolving and that a new phase is dawning.

References

Anonymous (1876) Objects of the London Epidemiological Society. In Transaction of the London Epidemiological Society. Vol. III. Sessions 1866–1876. London: Hardwicke and Bogue, pp 3–7

Bartley H (1832) Illustrations of cholera asphyxia in its different stages selected from cases treated at the Cholera Hospital Rivington Street. S.H. Jackson, New York

Berkson J (1946) Limitations of the application of four-fold table analysis to hospital data. Biometrics 2:47–53

Berkson J (1955) The statistical study of association between smoking and lung cancer. Mayo Clin Proc 30:319–348

Blackburn H, Labarthe D (2012) Stories from the evolution of guidelines for causal inference in epidemiologic associations: 1953-1965. Am J Epidemiol 176:1071–1077

Bracken M (2003) The first epidemiologic text. Am J Epidemiol 157:855–856

Buck C (1975) Popper's philosophy for epidemiologists. Int J Epidemiol 4:159–168

Carter KC (1983) Ignaz Semmelweis. The etiology, concept and prophylaxis of childbed fever. University of Wisconsin Press, Wisconsin

Chalmers I (2002) MRC Therapeutic Trials Committee's report on serum treatment of lobar pneumonia, BMJ 1934. JLL Bulletin: Commentaries on the history of treatment evaluation (www.jameslindlibrary.org)

Chalmers I (2010) Joseph Asbury Bell and the birth of randomized trials. JLL Bulletin: Commentaries on the history of treatment evaluation (www.jameslindlibrary.org)

Cole SR, Hernan MA, Robins JM, Anastos K, Chmiel J, Detels R, Ervin C, Feldman J, Greenblatt R, Kingsley L, Lai S, Young M, Cohen M, Munoz A (2003) Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. Am J Epidemiol 158:687–694

Comstock GW (2004) Cohort analysis: W.H. Frost's contributions to the epidemiology of tuberculosis and chronic disease. In: Morabia A (ed) History of epidemiological methods and concepts. Birkhäuser, Basel, pp 223–231

Copas J (1973) Randomization models for the matched and unmatched 2 × 2 tables. Biometrika 60:467–476

Copeland KT, Checkoway H, McMichael AJ, Holbrook RH (1977) Bias due to misclassification in the estimation of relative risk. Am J Epidemiol 105:488–495

Cornfield J (1951) A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. J Natl Cancer Inst 11:1269–1275

Creighton C (1894) A history of epidemics in Britain. Cambridge University Press, London

Darroch J (1997) Biologic synergism and parallelism. Am J Epidemiol 145:661–668

Diamond E, Lilienfeld A (1962a) Effects of errors in classification and diagnosis in various types of epidemiological-studies. Am J Public Health Nations Health 52:1137–1144

Diamond E, Lilienfeld A (1962b) Misclassification errors in 2 × 2 tables with one margin fixed: some further comments. Am J Public Health Nations Health 52:2106–2110

Doll R (1991) Prospective or retrospective: what's in a name? BMJ 302:528

Doll R, Hill AB (1950) Smoking and carcinoma of the lung; preliminary report. Br Med J 2:739–748

Doll R, Hill AB (1954) The mortality of doctors in relation to their smoking habits: a preliminary report. Br Med J 1:1451–1455

Dorn HF (1959) Tobacco consumption and mortality from cancer and other diseases. Public Health Rep 74:581–593

Elandt-Johnson RC (1975) Definition of rates: some remarks on their use and misuse. Am J Epidemiol 102:267–271

English JP, Willius FA, Berkson J (1940) Tobacco and coronary disease. J Am Med Assoc 115:1327–1328

Farr W (1838) “On prognosis” by William Farr (British Medical Almanack 1838; Supplement 199–216) Part 1 (pp 199–208). In: Morabia A (ed) History of epidemiological methods and concepts (2004). Birkhäuser, Basel, pp 159–178

Frost WH (1933) Risk of persons in familial contact with tuberculosis. Am J Public Health 23:426–432

Frost WH (1941) Epidemiology. In: Papers of Wade Hampton Frost, M.D.: a contribution to epidemiological methods. Common Wealth Fund, New York

Glendy RE, Levine SA, White PD (1937) Coronary disease in youth. J Am Med Assoc 109:1775–1781

Goldberger J, Wheeler GA, Sydenstricker E (1920) A study of the relation of family Income and other economic factors to pellagra incidence in seven cotton-mill villages of South Carolina in 1916. Public Health Rep 35:2673–2714

Graunt J (1662) Natural and political observations made upon the bills of mortality. www.ac.wwu.edu/∼stephan/Graunt/. Johns Hopkins University Press (1939), Baltimore. 8-4-2004

Greenland S, Brumback B (2002) An overview of relations among causal modelling methods. Int J Epidemiol 31:1030–1037

Greenland S, Neutra R (1980) Control of confounding in the assessment of medical technology. Int J Epidemiol 9:361–367

Greenland S, Robins JM (1986) Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 15:413–419

Greenland S, Robins JM (2009) Identifiability, exchangeability and confounding revisited. Epidemiol Perspect Innov 6:4

Greenland S, Thomas DC (1982) On the need for the rare disease assumption in case-control studies. Am J Epidemiol 116:547–553

Greenland S, Thomas DC, Morgenstern H (1986) The rare-disease assumption revisited. A critique of “estimators of relative risk for case-control studies”. Am J Epidemiol 124:869–883

Greenland S, Pearl J, Robins JM (1999) Causal diagrams for epidemiologic research. Epidemiology 10:37–48

Greenwood M (1935) Epidemics & crowd diseases: introduction to the study of epidemiology. Ayer Company Publishers, Incorporated, North Stratford

Hammond EC (1955) Causes and effect. In: Wynder EL (ed) The biologic effects of tobacco, with emphasis on the clinical and experimental aspects. Little, Brown, Boston, pp 171–196

Hammond EC, Horn D (1958a) Smoking and death rates; report on forty-four months of follow-up of 187,783 men. I. Total mortality. J Am Med Assoc 166:1159–1172

Hammond EC, Horn D (1958b) Smoking and death rates; report on forty-four months of follow-up of 187,783 men. II. Death rates by cause. J Am Med Assoc 166:1294–1308

Hammond EC, Selikoff IJ, Seidman H (1979) Asbestos exposure, cigarette smoking and death rates. Ann N Y Acad Sci 330:473–490

Hardy A (1999) Food, hygiene, and the laboratory: a short history of food poisoning in Britain, circa 1850–1950. Soc Hist Med 12:293–311

Hernan MA, Robins JM (2006) Estimating causal effects from epidemiological data. J Epidemiol Community Health 60:578–586

Hernan MA, Brumback B, Robins JM (2000) Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 11:561–570

Hernan MA, Hernandez-Diaz S, Robins JM (2004) A structural approach to selection bias. Epidemiology 15:615–625

Hill AB (1939) Principles of medical statistics. The Lancet Ltd, London

Hill AB (1965) Environment and disease: association or causation? Proc R Soc Med 58:295–300

Hoffman FL (1931) Cancer and smoking habits. Ann Surg 93:50-67

Hume D (1978) A treatise of human nature (1739). Oxford University Press, Oxford

Johnson S (2006) The ghost map: the story of London's most terrifying epidemic – and how it changed science, cities and the modern world. Riverheads Books, New York

Johnson WM (1929) Tobacco smoking. A clinical study. J Am Med Assoc 93:665–667

Keys A, Kihlberg JK (1963) Effect of misclassification on estimated relative prevalence of a characteristic. I. Two populations infallibly distinguished. II. Errors in two variables. Am J Public Health Nations Health 53:1656–1665

Kish L (1959) Some statistical problems in research design. Am Sociol Rev 24:328–338

Kleinbaum DG, Kupper LL, Morgenstern H (1982) Epidemiologic research: principles & quantitative methods. Van Nostrand Reinhold, New York

Kuh D, Ben-Shlomo Y (2004) A life course approach to chronic disease epidemiology. Oxford University Press, New York

Kupper LL, Karon JM, Kleinbaum DG, Morgenstern H, Lewis DK (1981) Matching in epidemiologic studies: validity and efficiency considerations. Biometrics 37:271–291

Lane-Claypon J (1926) A further report on cancer of the breast: reports on public health and medical subjects. Ministry of Health 32, 1–189. His Majesty's Stationary Office, London

Last JMe (2001) A dictionary of epidemiology. Oxford University Press, Oxford

Levin ML, Goldstein H, Gerhardt PR (1950) Cancer and tobacco smoking: a preliminary report. J Am Med Assoc 143:336–338

Lilienfeld AM (1976) Foundations of epidemiology. Oxford University Press, New York

Lilienfeld AM, Graham S (1958) Validity of determining circumcision status by questionnaire as related to epidemiological studies of cancer of the cervix. J Natl Cancer Inst 21:713–720

Lind J (1753) A treatise of scurvy. University Press (1953), Edinburgh

Lombard HL, Doering CR (1928) Cancer studies in Massachusetts. 2. Habits, characteristics and environment of individuals with and without cancer. N Engl J Med 195:481–487

Louis PCA (1836) Researches on the effects of bloodletting in some inflammatory diseases. Hilliard, Gray and Company, Boston

MacMahon B, Pugh TF (1970) Epidemiology – principles and methods. Little, Brown and Co, Boston

MacMahon B, Pugh TF, Ipsen J (1960) Epidemiologic methods. Little, Brown and Co, Boston

Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 22:719–748

Marmor M, Friedman-Kien AE, Laubenstein L, Byrum RD, William DC, D'onofrio S, Dubin N (1982) Risk factors for Kaposi's sarcoma in homosexual men. Lancet 1:1083–1087

McMichael AJ (1999) Prisoners of the proximate: loosening the constraints on epidemiology in an age of change. Am J Epidemiol 149:887–897

McNeill WH (1977) Plagues and people. Anchor Books, New York

Medical Research Council (1952) Reports for the years 1948–1950. HMSO, London

Miettinen OS (1976) Estimability and estimation in case-referent studies. Am J Epidemiol 103:226–235

Miettinen OS (1985) Theoretical epidemiology. Wiley, New York

Miettinen OS, Cook EF (1981) Confounding: essence and detection. Am J Epidemiol 114:593–603

Moore PS, Chang Y (1995) Detection of herpesvirus-like DNA sequences in Kaposi's sarcoma in patients with and without HIV infection. N Engl J Med 332:1181–1185

Morabia A (1991) On the origin of Hill's causal criteria. Epidemiology 2:367–369

Morabia A (1995) Poppers, Kaposi's sarcoma, and HIV infection: empirical example of a strong confounding effect? Prev Med 24:90–95

Morabia A (1996) P. C. A. Louis and the birth of clinical epidemiology. J Clin Epidemiol 49: 1327–1333

Morabia A (1998) Epidemiology and bacteriology in 1900: who is the handmaid of whom? J Epidemiol Community Health 52:617–618

Morabia A (2004) Epidemiology: an epistemological perspective. In: Morabia A (ed) History of epidemiological methods and concepts. Birkhäuser, Basel, pp 1–126

Morabia A (2007a) Epidemiologic interactions, complexity, and the lonesome death of Max von Pettenkofer. Am J Epidemiol 166:1233–1238

Morabia A (2007b) Epidemiological methods and concepts in the nineteenth century and their influences on the twentieth century. In: Holland WW, Olsen J, Florey C (eds) The development of modern epidemiology. Personal reports from those who were there. Oxford University Press, New York, pp 17–29

Morabia A (2008) Joseph Goldberger's research on the prevention of pellagra. J R Soc Med 101:566–568

Morabia A (2009) Epidemic and population patterns in the Chinese Empire (243 B.C.E. to 1911 C.E.): quantitative analysis of a unique but neglected epidemic catalogue. Epidemiol Infect 137:1361–1368

Morabia A (2010) Janet Lane-Claypon – interphase epitome. Epidemiology 21:573–576

Morabia A (2011) History of the modern epidemiological concept of confounding. J Epidemiol Commun H 65:297–300

Morabia A (2012) Quality, originality, and significance of the 1939 Tobacco consumption and lung carcinoma article by Mueller, including translation of a section of the paper. Prev Med 55:171–177

Morabia A (2013a) Epidemiology's 350th Anniversary: 1662-2012. Epidemiology 24:179–183

Morabia A (2013b) "Lung cancer and tobacco consumption': technical evaluation of the 1943 paper by Schairer and Schoeniger published in Nazi Germany. J Epidemiol Commun H 67: 208–212

Morabia A, Guthold R (2007) Wilhelm Weinberg's 1913 large retrospective cohort study: a rediscovery. Am J Epidemiol 165:727–733

Morabia A, Hardy A (2005) The pioneering use of a questionnaire to investigate a food borne disease outbreak in early 20th century Britain. J Epidemiol Community Health 59:94–99

Morabia A, Rubenstein B, Victora CG (2013) Epidemiology and Public Health in 1906 England: Arthur Newsholmes Methodological Innovation to Study Breastfeeding and Fatal Diarrhea. Am J Public Health 103:in press

Morgenstern H, Kleinbaum DG, Kupper LL (1980) Measures of disease incidence used in epidemiologic research. Int J Epidemiol 9:97–104

Morris JN (1957) Uses of epidemiology. E. & S. Livingstone, Edinburgh

Mueller F (1939a) Abuse of tobacco and carcinoma of the lung. JAMA 113:1372

Mueller F (1939b) Tabakmissbrauch und Lungenkarcinoma. Z Krebsforsch 49:57–85

Muñoz N, Bosch FX, de SS, Tafur L, Izarzugaza I, Gili M, Viladiu P, Navarro C, Martos C, Ascunce N (1992) The Causal Link Between Human Papillomavirus and Invasive Cervical Cancer: A Population-Based Case-Control Study in Colombia and Spain. Int J Cancer 52:743–749

Ochsner A, DeBakey M (1939) Primary pulmonary malignancy: treatment by total pneumonectomy. Analysis of 79 collected cases and presentation of 7 personal cases. Surg Gynecol Obstet 68:435–451

Ochsner A, DeBakey ME, Dixon L (1947) Primary pulmonary malignancy treated by resection: an analysis of 129 cases. Ann Surg 125:522–540

Newell DJ (1962) Errors in the interpretation of errors in epidemiology. Am J Public Health 52:1925–1928

Pearl R (1929a) A note on the association of diseases. Science 70:191–192

Pearl R (1929b) Cancer and tuberculosis. Am J Hyg 9:97–159

Pearl R (1938) Tobacco smoking and longevity. Science 87:216–217

Pearl J (1995) Causal diagrams for empirical research. Biometrika 82:669–710

Pearl J (2000) Causality: models, reasoning and inference. Cambridge University Press, Cambridge

Petty W (1927) Concerning the study of the art of medicine. In: Lansdowne M (ed) The Petty papers. Constable, London, pp 168–180

Piaget J (1970) Genetic epistemology. Columbia University Press, New York

Porter R (1997) The greatest benefit to mankind: a medical history of humanity. Harper Collins, London

Press DJ, Pharoah P (2010) Risk factors for breast cancer: a re-analysis of two case-control studies from 1926 and 1931. Epidemiology 21:566–572

Proctor RN (2006) Angel H Roffo: the forgotten father of experimental tobacco carcinogenesis. Bull World Health Organ 84:494–496

Rothman KJ (1974) Synergy and antagonism in cause-effect relationships. Am J Epidemiol 99:385–388

Rothman KJ (1976) Causes. Am J Epidemiol 104:587–592

Rothman KJ (1977) Epidemiologic methods in clinical trials. Cancer 39:1771–1775

Rothman KJ (1986) Modern epidemiology. Little Brown and Co, Boston

Rothman KJ (ed) (1988) Causal inference. Epidemiology Resources, Boston

Rothman KJ, Greenland S (1998) Modern epidemiology. Lipincott Williams Wilkins, Philadelphia

Rothman KJ, Greenland S, Lash TL (2008) Modern epidemiology. Lipincott Williams Wilkins, Philadelphia

Roberts RS, Spitzer WO, Delmore T, Sackett DL (1978) An empirical demonstration of Berkson's bias. J Chronic Dis 31:119–128

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized treatments. J Educ Psychol 66:688–701

Rubin DB (1990) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Comment: Neyman (1923) and causal inference in experiments and observational studies. Stat Sci 5:472–480

Sackett DL (1979) Bias in analytic research. J Chronic Dis 32:51–63

Schairer E, Schoeniger E (1943) Lungenkrebs und Tabakverbrauch. Z Krebsforsch 54:261–269

Schairer E, Schoeniger E (2001) Lung cancer and tobacco consumption. Int J Epidemiol 30:24–27

Schrek R, Baker LA, Ballard GP, Dolgoff S (1950) Tobacco smoking as an etiologic factor in disease. I. Cancer. Cancer Res 10:49–58

Schwartz S, Susser E (2006) Commentary: what can epidemiology accomplish? Int J Epidemiol 35:587–590

Simpson EH (1951) The interpretation of interaction in contingency tables. J R Stat Soc Ser B 13:238–241

Snow J (1855) On the mode of communication of cholera. Churchill, London

Snow J (1936) Snow on cholera. The Commonwealth Fund, New York

Splawa-Neyman J (1990) On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (1923). Stat Sci 5:465–480

Stocks P, Karn M (1933) A co-operative study of the habits, home life, dietary and family histories of 450 cancer patients and of an equal number of control patients. Ann Eugen 5:237–279

Susser M (1973) Causal thinking in the health sciences. Oxford, New York

Susser M (1977) Judgement and causal inference: criteria in epidemiologic studies. Am J Epidemiol 105:1–15

Susser M, Susser E (1996) Choosing a future for epidemiology: I. Eras and paradigms. Am J Public Health 86:668–673

Susser E, Schwartz S, Morabia A, Bromet E (2006) Psychiatric epidemiology. Oxford University Press, Oxford

Szklo M, Nieto F (2000) Epidemiology: beyond the basics. Aspen, Gaithersburg

Templeton AR (2000) Epistasis and complex traits. In: Wolf J, Brodie E, Wade M (eds) Epistasis and the evolutionary process. Oxford University Press, New York, pp 41–57

Therapeutic Trial Committee of the Medical Research Council (1934) The serum treatment of lobar pneumonia. Lancet 1:290–295

The Surgeon General's Advisory Committee (1964) Smoking and health. Public Health Service Publication No 1103, Washington

Troehler U (2000) 'To improve the evidence of medicine': the 18th century British origins of a critical approach. Royal College of Physicians, Edinburgh

Vandenbroucke JP (1991) Prospective or retrospective: what's in a name? BMJ 302:249–250

Vandenbroucke JP (2004) The history of confounding. In: Morabia A (ed) History of epidemiological methods and concepts. Birkhäuser, Basel, pp 313–326

Vanderweele TJ, Robins JM (2007) Four types of effect modification: a classification based on directed acyclic graphs. Epidemiology 18:561–568

Vandenbroucke JP, Eelkman Rooda HM, Beukers H (1991) Who made John Snow a hero? Am J Epidemiol 133:967–973

Villalba Jd (1803) Epidemiologia española. Fermi Villalpando, Madrid

Vinten-Johansen P, Brody H, Paneth N, Rachman S, Rip MR (2003) Cholera, chloroform and the science of medicine: a life of John Snow. Oxford University Press, Oxford

von Pettenkofer M (1869) Boden und Grundwasser in ihren Beziehunen zu Cholera und Typhus. Zeitschrift für Biologie 5:170–310

Wainwright J (1931) A comparison of conditions associated with breast cancer in great Britain and America. Am J Cancer 15:2610–2645

Weinberg W (1913) Die Kinder der Tuberkulosen. S Hirzel, Leipzig

Winkelstein W Jr (2004) Vignettes of the history of epidemiology: three firsts by Janet Elizabeth Lane-Claypon. Am J Epidemiol 160:97–101

Wynder E, Graham EA (1950) Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma: a study of six hundred and eighty four proved cases. J Am Med Assoc 143: 329–336

Wynder EL, Cornfield J, Schroff P, Doraiswani K (1954) A study of environmental factors in carcinoma of the cervix. Am J Obstet Gynecol 68:1016–1047

Yang Q, Khoury MJ, Flanders WD (1997) Sample size requirements in case-only designs to detect gene-environment interaction. Am J Epidemiol 146:713–720

Yule GU (1903) Notes on the theory of association of attributes in statistics. Biometrika 2:121–134

Zhang FF, Michaels DC, Mathema B, Kauchali S, Chatterjee A, Ferris DC, James TM, Knight J, Dounel M, Tawfik HO, Frohlich JA, Kuang L, Hoskin EK, Veldman FJ, Baldi G, Mlisana KP, Mametja LD, Diaz A, Khan NL, Sternfels P, Sevigny JJ, Shamam A, Morabia A (2004) Evolution of epidemiologic methods and concepts in selected textbooks of the 20th century. In: Morabia A (ed) History of epidemiological methods and concepts. Birkhauser, Basel, pp 351–362

Basic Concepts

3

Kenneth J. Rothman and Sander Greenland

Contents

K.J. Rothman (✉)
RTI Health Solutions, Research Triangle Institute, Research Triangle Park, NC, USA

S. Greenland
Department of Epidemiology, School of Public Health, University of California, Los Angeles, CA, USA

Department of Statistics, UCLA College of Letters and Science, Los Angeles, CA, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_44,

3.1 Introduction

Epidemiology is the science that focuses on the occurrence of disease in its broadest sense, with the fundamental aim to understand and to control its causes. This chapter deals with the conceptual building blocks of epidemiology. First, we offer a model for causation, from which a variety of insights relevant to epidemiological understanding emerge. We then discuss the basis by which we attempt to infer that an identified factor is indeed a cause of disease; the guidelines lead us through a rapid review of modern scientific philosophy. The remainder of the chapter deals with epidemiological fundamentals of measurement, including the measurement of disease and the measurement of causal effects. More detailed discussion can be found in many textbooks; the present material follows the viewpoints, terminology, and development in Modern Epidemiology, 3rd edition (Rothman et al. 2008).

3.2 Causation and Causal Inference

3.2.1 A General Model of Causation

While useful in everyday life, ordinary concepts of causation are too primitive to serve well as the basis for scientific theories. This shortcoming may be especially true in the health and social sciences, in which typical causes are neither necessary nor sufficient to bring about effects of interest. Hence, as has long been recognized, there is a need to develop a more refined conceptual model that can serve as a starting point in discussions of causation. In particular, such a model should facilitate problems of multifactorial causation, confounding, interdependence of effects, direct and indirect effects, levels of causation, and systems or webs of causation (MacMahon and Pugh 1967, 1970; Susser 1973). This section describes one starting point, the sufficient-component cause model (or sufficient-cause model), which has proven useful in elucidating certain concepts in individual mechanisms of causation. Later, we will discuss causal models better suited for population studies of causation in the absence of mechanistic theories.

To begin, we need to define *cause*. For our purposes, we can define a cause of a specific disease event as an antecedent event, condition, or characteristic that was necessary for the occurrence of a specific instance of the disease at the moment it occurred, given that other conditions are fixed. In other words, a cause of a disease event is an event, condition, or characteristic that preceded the disease event and without which the disease event either would not have occurred at all or would not have occurred until some later time. With this definition, it may be that no specific event, condition, or characteristic is sufficient by itself to produce disease. This definition, then, does not define a complete causal mechanism but only a component of it.

A common characteristic of the concept of causation that we develop early in life is the assumption of a one-to-one correspondence between the observed cause and effect. Each cause is seen as necessary *and* sufficient in itself to produce the effect. Thus, the flick of a light switch appears to be the singular cause that makes the lights go on. There are less evident causes, however, that also operate to produce the effect: The need for an unspent bulb in the light fixture, wiring from the switch to the bulb, and voltage to produce a current when the circuit is closed. To achieve the effect of turning on the light, each of these is equally as important as moving the switch because absence of any of these components of the causal constellation will prevent the effect.

For many people, the roots of early causal thinking persist and become manifest in attempts to find single causes as explanations for observed phenomena. But experience and reflection should easily persuade us that the cause of any effect must consist of a constellation of components that act in concert (Mill 1843). A "sufficient cause," which means a complete causal mechanism, can be defined as a set of minimal conditions and events that inevitably produce disease; "minimal" implies that all of the conditions or events are necessary (Mackie 1965). In disease etiology, the completion of a sufficient cause may be considered equivalent to the onset of disease. (Onset here refers to the onset of the earliest stage of the disease process, rather than the onset of signs or symptoms.) For biological effects, most and sometimes all of the components of a sufficient cause are unknown (Rothman 1976).

For example, tobacco smoking is a cause of lung cancer, but by itself, it is not a sufficient cause. First, the term *smoking* is too imprecise to be used in a causal description. One must specify the type of smoke (e.g., cigarette, cigar, pipe), whether it is filtered or unfiltered, the manner and frequency of inhalation, and the onset and duration of smoking. More important, smoking, even defined explicitly, will not cause cancer in everyone. So who are those who are "susceptible" to the effects of smoking? Or, to put it in other terms, what are the other components of the causal constellation that act with smoking to produce lung cancer?

When causal components remain unknown, we are inclined to assign an equal risk to all individuals whose causal status for some components is known and identical. Thus, men who are heavy cigarette smokers are said to have approximately a 10% lifetime risk of developing lung cancer. Some interpret this statement to mean that all men would be subject to a 10% probability of lung cancer if they were to become heavy smokers, as if the outcome, aside from smoking, were purely a matter of chance. In contrast, we view the assignment of equal risks as reflecting nothing more than assigning to everyone within a specific category, in this case male heavy smokers, the average of the individual risks for people in that category. In the classical view, these risks are either one or zero, according to whether the individual will or will not get lung cancer.

We cannot measure the individual risks, and assigning the average value to everyone in the category reflects nothing more than our ignorance about the determinants of lung cancer that interact with cigarette smoke. It is apparent from epidemiological data that some people can engage in chain smoking for many decades without developing lung cancer. Others are or will become "primed" by unknown circumstances and need only to add cigarette smoke to the nearly sufficient constellation of causes to initiate lung cancer. In our ignorance of these hidden

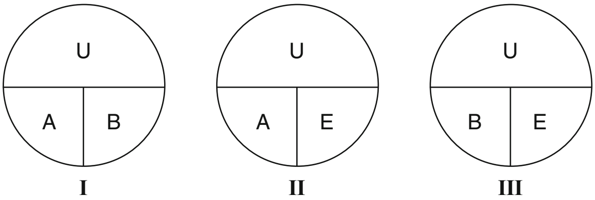


Fig. 3.1 Three sufficient causes of a disease

causal components, the best we can do in assessing risk is to classify people according to measured causal risk indicators and then assign the average risk observed within a class to persons within the class. As knowledge expands, the risk estimates assigned to people will depart from average according to the presence or absence of other factors that affect the risk.

For example, we now know that smokers with substantial asbestos exposure are at higher risk of lung cancer than those who lack asbestos exposure. Consequently, with adequate data, we could assign different risks to heavy smokers based on their asbestos exposure. Within categories of asbestos exposure, the average risks would be assigned to all heavy smokers until other risk factors are identified.

Figure 3.1 provides a schematic diagram of sufficient causes in a hypothetical individual. Each constellation of component causes represented in Fig. 3.1 is minimally sufficient to produce the disease; that is, there is no redundant or extraneous component cause – each one is a necessary part of that specific causal mechanism. A specific component cause may play a role in one, two, or all three of the causal mechanisms pictured.

Figure 3.1 does not depict aspects of the causal process such as prevention, sequence, or timing of action of the component causes, dose, and other complexities. These aspects of the causal process must be accommodated in the model by an appropriate definition of each causal component. Thus, if the disease is lung cancer and the factor E represents cigarette smoking, it might be defined more explicitly as smoking at least two packs a day of unfiltered cigarettes for at least 20 years. If the outcome is smallpox, which is completely prevented by immunization, U could represent "unimmunized." More generally, preventive effects of a factor C can be represented by placing its complement "no C" within sufficient causes.

3.2.2 Strength of Effects

We will call the set of conditions necessary and sufficient for a factor to produce disease the *causal complement* of the factor. Thus, the condition "(A and U) or (B and U)" is the causal complement of E in the above example. The strength of a factor's effect on a population depends on the relative prevalence of its causal complement. This dependence of the effects of a specific component cause on the prevalence of its causal complement has nothing to do with the biological mechanism of the component's action, since the component is an equal partner in each mechanism in which it appears. Nevertheless, a factor will appear to have a strong effect if its causal complement is common. Conversely, a factor with a rare causal complement will appear to have a weak effect.

In epidemiology, the strength of a factor's effect is usually measured by the change in disease frequency produced by introducing the factor into a population. This change may be measured in absolute or relative terms. In either case, the strength of an effect may have tremendous public health significance, but it may have little biological significance. The reason is that given a specific causal mechanism, *any* of the component causes can have strong or weak effects. The actual identity of the constituent components of the cause amount to the biology of causation, whereas the strength of a factor's effect depends on the time-specific distribution of its causal complements in the population. Over a span of time, the strength of the effect of a given factor on the occurrence of a given disease may change, because the prevalence of its causal complements in various mechanisms may also change. The causal mechanisms in which the factor and its complements act could remain unchanged, however.

3.2.3 Interaction Among Causes

Two component causes acting in the same sufficient cause may be thought of as interacting biologically to produce disease. Indeed, one may define biological interaction as the participation of two component causes in the same sufficient cause. Such interaction is also known as causal coaction or joint action. The joint action of the two component causes does not have to be simultaneous action: One component cause could act many years before the other, but it would have to leave some effect that interacts with the later component.

For example, suppose a traumatic injury to the head leads to a permanent disturbance in equilibrium. Many years later, the faulty equilibrium may lead to a fall while walking on an icy path, causing a broken hip. The causal mechanism for the broken hip includes the traumatic injury to the head as a component cause, along with its consequence of a disturbed equilibrium. The causal mechanism also includes the walk along the icy path. These two component causes have interacted with one another, although their time of action is many years apart. They also would interact with the other component causes, such as the type of footwear, the absence of a handhold, and any other conditions that were necessary to the causal mechanism of the fall and the broken hip that resulted.

The degree of observable interaction between two specific component causes depends on how many different sufficient causes produce disease, and the proportion of cases that occur through sufficient causes in which the two component causes both play some role. For example, in Fig. 3.2, suppose that G were only a hypothetical substance that did not actually exist. Consequently, no disease would occur from sufficient cause II, because it depends on an action by G, and factors B and F would act only through the distinct mechanisms represented by sufficient causes I and III. Thus, B and F would be biologically independent. Now suppose G is present; then B and F would interact biologically. Furthermore, if C is completely

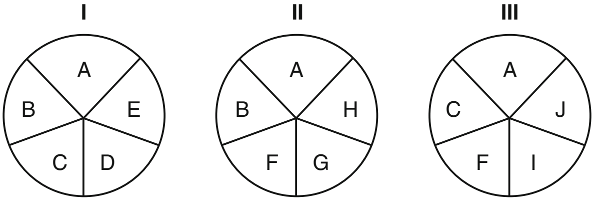


Fig. 3.2 Another example of three sufficient causes of a disease

absent, then cases will occur only when factors B and F act together in the mechanism represented by sufficient cause II. Thus, the extent or apparent strength of biological interaction between two factors is dependent on the prevalence of other factors.

3.2.4 Proportion of Disease Due to Specific Causes

In Fig. 3.1, assuming that the three sufficient causes in the diagram are the only ones operating, what fraction of disease is caused by U? The answer is all of it; without U, there is no disease. U is considered a "necessary cause." What fraction is due to E? E causes disease through two mechanisms, I and II, and all disease arising through either of these two mechanisms is due to E. This is not to say that all disease is due to U alone or that a fraction of disease is due to E alone; no component cause acts alone. Rather, these factors interact with their complementary factors to produce disease.

There is a tendency to think that the sum of the fractions of disease attributable to each of the causes of the disease should be 100%. For example, in their widely cited work, *The Causes of Cancer*, Doll and Peto created a table giving their estimates of the fraction of all cancers caused by various agents; the total for the fractions was nearly 100% (Doll and Peto 1981, Table 20). Although they acknowledged that any case could be caused by more than one agent, which would mean that the attributable fractions would not sum to 100%, they referred to this situation as a "difficulty" and an "anomaly." It is, however, neither a difficulty nor an anomaly but simply a consequence of allowing for the fact that no event has a single agent as the cause. The fraction of disease that can be attributed to each of the causes of disease in all the causal mechanisms actually has no upper limit: For cancer, or any disease, the upper limit for the total of the fraction of disease attributable to all the component causes of all the causal mechanisms that produce it is not 100% but infinity. Only the fraction of disease attributable to a single component cause cannot exceed 100%.

3.2.5 Induction Period and Latent Period

The diagram of causes in Fig. 3.2 also provides a model for conceptualizing the *induction period,* which may be defined as the period of time from causal action until disease occurrence. If, in sufficient cause I, the sequence of action of the causes is A, B, C, D, and E, and we are studying the effect of B, which, let us assume, acts at a narrowly defined point in time, disease does not occur immediately after B acts. It occurs only after the sequence is completed, so there will be a delay while C, D, and finally E act. When E acts, disease occurs. The interval between the action of B and the disease occurrence is the induction time for the effect of B.

In the example given earlier of an equilibrium disorder leading to a later fall and hip injury, the induction time between the occurrence of the equilibrium disorder and the later hip injury might be very long. In an individual instance, we would not know the exact length of an induction period, since we cannot be sure of the causal mechanism that produces disease in an individual instance, nor when all the relevant component causes acted. We can characterize the induction period relating the action of a component cause to the occurrence of disease in general, however, by accumulating data for many individuals. A clear example of a lengthy induction time is the cause-effect relation between exposure of a female fetus to diethylstilbestrol (DES) and the subsequent development of clear cell adenocarcinoma of the vagina. The cancer usually occurs between the ages of 15 and 30. Since the causal exposure to DES occurs early in pregnancy, there is an induction time of about 15–30 years for the carcinogenic action of DES. During this time, other causes presumably are operating; some evidence suggests that hormonal action during adolescence may be part of the mechanism (Rothman 1981).

It is incorrect to characterize a disease itself as having a lengthy or brief induction time. The induction time can be conceptualized only in relation to a specific component cause. Thus, we say that the induction time relating DES to clear cell carcinoma of the vagina is 15–30 years, but we cannot say that 15–30 years is the induction time for clear cell carcinoma in general. Since each component cause in any causal mechanism can act at a time different from the other component causes, each can have its own induction time. For the component cause that acts last, the induction time equals zero. If another component cause of clear cell carcinoma of the vagina that acts during adolescence were identified, it would have a much shorter induction time for its carcinogenic action than DES. Thus, induction time characterizes a specific cause-effect pair rather than just the effect.

Disease, once initiated, will not necessarily be apparent. The time interval between disease occurrence and detection has been termed the *latent* period (Rothman 1981), although others have used this term interchangeably with induction period. The latent period can be reduced by improved methods of disease detection. The induction period, on the other hand, cannot be reduced by early detection of disease, since disease occurrence marks the end of the induction period. Earlier detection of disease, however, may reduce the apparent induction period (the time between causal action and disease detection), since the time when disease is detected, as a practical matter, is usually used to mark the time of disease occurrence. Thus, diseases such as slow-growing cancers may appear to have long induction periods with respect to many causes because they have long latent periods. The latent period, unlike the induction period, is a characteristic of the disease and the detection effort applied to the person with the disease.

Although it is not possible to reduce the induction period proper by earlier detection of disease, it may be possible to observe intermediate stages of a causal mechanism. The increased interest in biomarkers such as DNA adducts is an example of attempting to focus on causes more proximal to the disease occurrence. Such biomarkers may reflect the effects of earlier acting agents on the organism.

Some agents may have a causal action by shortening the induction time of other agents. Suppose that exposure to factor A leads to epilepsy after an interval of 10 years, on the average. It may be that exposure to a drug, B, would shorten this interval to 2 years. Is B acting as a catalyst or as a cause of epilepsy? The answer is both: A catalyst *is* a cause. Without B, the occurrence of epilepsy comes 8 years later than it comes with B, so we can say that B causes the onset of the early epilepsy. It is not sufficient to argue that the epilepsy would have occurred anyway. First, it would not have occurred at that time, and the time of occurrence is part of our definition of an event. Second, epilepsy will occur later only if the individual survives an additional 8 years, which is not certain. Agent B not only determines when the epilepsy occurs, it can determine whether it occurs. Thus, we should call any agent that acts as a catalyst of a causal mechanism, speeding up an induction period for other agents, as a cause in its own right. Similarly, any agent that postpones the onset of an event, drawing out the induction period for another agent, is a preventive. It should not be too surprising to equate postponement to prevention: We routinely use such an equation when we employ the euphemism that we prevent death, which actually can only be postponed. What we prevent is death at a given time, in favor of death at a later time.

3.2.6 Philosophy of Scientific Inference

Modern science began to emerge around the sixteenth and seventeenth centuries, when the knowledge demands of emerging technologies (such as artillery and transoceanic navigation) stimulated inquiry into the origins of knowledge. An early codification of the scientific method was Francis Bacon's *Novum Organum,* published in 1620, which presented an *inductivist* view of science. In this philosophy, scientific reasoning is said to depend on making generalizations, or inductions, from observations to general laws of nature; the observations are said to induce the formulation of a natural law in the mind of the scientist. Thus, an inductivist would have said that Jenner's observation of lack of smallpox among milkmaids induced in Jenner's mind the theory that cowpox (common among milkmaids) conferred immunity to smallpox. Inductivist philosophy reached a pinnacle of sorts in the canons of John Stuart Mill (1843), which evolved into inferential criteria that are still in use today.

Inductivist philosophy was a great step forward from the medieval scholasticism that preceded it, for at least it demanded that a scientist make careful observations of people and nature, rather than appeal to faith, ancient texts, or authorities. Nonetheless, by the eighteenth century, the Scottish philosopher David Hume had described a disturbing deficiency in inductivism: An inductive argument carried no logical force; instead, such an argument represented nothing more than an *assumption* that certain events would in the future follow in the same pattern as they had in the past. Thus, to argue that cowpox caused immunity to smallpox because no one got smallpox after having cowpox corresponded to an unjustified assumption that the pattern observed so far (no smallpox after cowpox) will continue into the future. Hume pointed out that, even for the most reasonable sounding of such assumptions, there was no logic or force of necessity behind the inductive argument.

Causal inference based on mere coincidence of events constitutes a logical fallacy known as *post hoc ergo propter hoc* (Latin for "after this therefore on account of this"). This fallacy is exemplified by the inference that the crowing of a rooster is necessary for the sun to rise because sunrise is always preceded by the crowing. The *post hoc* fallacy is a special case of a more general logical fallacy known as the *fallacy of affirming the consequent*. This fallacy of confirmation takes the following general form: "We know that if H is true, B must be true; and we know that B is true; therefore H must be true." This fallacy is used routinely by scientists in interpreting data. It is used, for example, when one argues as follows: "if sewer service causes heart disease, then heart disease rates should be highest where sewer service is available; heart disease rates are indeed highest where sewer service is available; therefore, sewer service causes heart disease." Here, H is the hypothesis "sewer service causes heart disease" and B is the observation "heart disease rates are highest where sewer service is available." The argument is of course logically unsound, as demonstrated by the fact that we can imagine many ways in which the premises could be true but the conclusion false; for example, economic development could lead to both sewer service and elevated heart disease rates, without any effect of the latter on the former.

3.2.7 Refutationism

Many philosophers and scientists from Hume's time forward attempted to set out a firm logical basis for scientific reasoning. In the 1920s, most notable among these was the school of logical positivists, who sought a logic for science that could lead inevitably to correct scientific conclusions, in much the way rigorous logic can lead inevitably to correct conclusions in mathematics. Other philosophers and scientists, however, had started to suspect that scientific hypotheses can never be proven or established as true in any logical sense. For example, a number of philosophers noted that scientific statements can only be found to be consistent with observation but cannot be proven or disproven in any "airtight" logical or mathematical sense (Duhem 1906, transl. 1954; Popper 1934, transl. 1959; Quine 1951). This fact is sometimes called the problem of *non-identification* or *underdetermination* of theories by observations (Curd and Cover 1998). In particular, available observations are always consistent with several hypotheses that themselves are mutually inconsistent, which explains why (as Hume noted) scientific theories cannot be logically proven. In particular, consistency between a hypothesis and observations is no proof of the hypothesis, because we can always invent alternative hypotheses that are just as consistent with the observations.

In contrast, a valid observation that is inconsistent with a hypothesis implies that the hypothesis as stated is false and so refutes the hypothesis. If you wring the rooster's neck before it crows and the sun still rises, you have disproved that the rooster's crowing is a necessary cause of sunrise. Or consider a hypothetical research program to learn the boiling point of water (Magee 1985). A scientist who boils water in an open flask and repeatedly measures the boiling point at 100°C will never, no matter how many confirmatory repetitions are involved, prove that 100°C is always the boiling point. On the other hand, merely one attempt to boil the water in a closed flask or at high altitude will refute the proposition that water always boils at 100°C.

According to Popper, science advances by a process of elimination that he called "conjecture and refutation." Scientists form hypotheses based on intuition, conjecture, and previous experience. Good scientists use deductive logic to infer predictions from the hypothesis and then compare observations with the predictions. Hypotheses whose predictions agree with observations are confirmed (Popper used the term "corroborated") only in the sense that they can continue to be used as explanations of natural phenomena. At any time, however, they may be refuted by further observations and might be replaced by other hypotheses that are more consistent with the observations. This view of scientific inference is sometimes called *refutationism* or *falsificationism.* Refutationists consider induction to be a psychologic crutch: Repeated observations did not in fact induce the formulation of a natural law, but only the belief that such a law has been found. For a refutationist, only the psychologic comfort provided by induction explains why it still has advocates.

One way to rescue the concept of induction from the stigma of pure delusion is to resurrect it as a psychologic phenomenon, as Hume and Popper claimed it was, but one that plays a legitimate role in hypothesis formation. The philosophy of conjecture and refutation places no constraints on the origin of conjectures. Even delusions are permitted as hypotheses, and therefore inductively inspired hypotheses, however psychological, are valid starting points for scientific evaluation. This concession does not admit a logical role for induction in confirming scientific hypotheses, but it allows the process of induction to play a part, along with imagination, in the scientific cycle of conjecture and refutation.

The philosophy of conjecture and refutation has profound implications for the methodology of science. The popular concept of a scientist doggedly assembling evidence to support a favorite thesis is objectionable from the standpoint of refutationist philosophy because it encourages scientists to consider their own pet theories as their intellectual property, to be confirmed, proven, and, when all the evidence is in, cast in stone and defended as natural law. Such attitudes hinder critical evaluation, interchange, and progress. The approach of conjecture and refutation, in contrast, encourages scientists to consider multiple hypotheses and to seek crucial tests that decide between competing hypotheses by falsifying one of them. Because falsification of one or more theories is the goal, there is incentive to depersonalize the theories. Criticism leveled at a theory need not be seen as criticism of the person who proposed it. It has been suggested that the reason why certain fields of science advance rapidly while others languish is that the rapidly advancing fields are propelled by scientists who are busy constructing and testing competing hypotheses; the other fields, in contrast, “are sick by comparison, because they have forgotten the necessity for alternative hypotheses and disproof” (Platt 1964).

The refutationist model of science has a number of valuable lessons for research conduct, especially of the need to seek alternative explanations for observations, rather than focus on the chimera of seeking scientific “proof” for some favored theory. Nonetheless, it is vulnerable to several criticisms (Pearce 1990), most notable among them being that observations (or some would say their interpretations) are themselves laden with theory (sometimes called the *Duhem-Quine thesis*; Curd and Cover 1998). Thus, observations can never provide the sort of definitive refutations that are the hallmark of popular accounts of refutationism. For example, there may be uncontrolled and even unimagined biases that have made our refutational observations invalid; to claim refutation is to assume as true the unprovable theory that no such bias exists. In other words, theories are underdetermined by observations, but refutations are theory-laden and thus also underdetermined. The net result is that logical certainty about either the truth or falsity of an internally consistent theory is impossible (Quine 1951).

3.2.8 Bayesianism

There is another philosophy of inference that, like refutationism, holds an objective view of scientific truth and a view of knowledge as tentative or uncertain but which focuses on evaluation of knowledge rather than truth. Like refutationism, the modern form of this philosophy evolved from the writings of eighteenth century British philosophers, but the focal arguments first appeared in a pivotal essay by Bayes (1763), and hence the philosophy is usually referred to as Bayesianism (Howson and Urbach 1993). Like refutationism, it did not reach a complete expression until after World War I, most notably in the writings of Ramsey (1931) and DeFinetti (1937), and, like refutationism, it did not begin to appear in epidemiology until the 1970s (see, e.g., Cornfield 1976).

The central problem addressed by Bayesianism is the following: In classical logic, a deductive argument can provide you no information about the truth or falsity of a scientific hypothesis unless you can be 100% certain about the truth of the premises of the argument. Consider the logical argument called *modus tollens*: “If H implies B, and B is false, then H must be false.” This argument is logically valid, but the conclusion follows only on the assumptions that the premises “H implies B” and “B is false” are true statements. If these premises are statements about the physical world, we cannot possibly know them to be correct with 100% certainty, since all observations are subject to error. Furthermore, the claim that “H implies B” will often depend on its own chain of deductions, each with its own premises of which we cannot be certain.

For example, if H is "television viewing causes homicides" and B is "homicide rates are highest where televisions are most common," the first premise used in modus tollens to test the hypothesis that television viewing causes homicides will be "If television viewing causes homicides, homicide rates are highest where televisions are most common." The validity of this premise is doubtful – after all, even if television does cause homicides, homicide rates may be low where televisions are common because of socioeconomic advantages in those areas.

Continuing to reason in this fashion, we could arrive at a more pessimistic state than even Hume imagined: Not only is induction without logical foundation, but *deduction* has no scientific utility because we cannot ensure the validity of all the premises. The Bayesian answer to this problem is partial, in that it makes a severe demand on the scientist and puts a severe limitation on the results. It says roughly this: If you can assign a degree of certainty, or personal probability, to the premises of your valid argument, you may use any and all the rules of probability theory to derive a certainty for the conclusion, and this certainty will be a logically valid consequence of your original certainties. The catch is that your concluding certainty, or *posterior probability*, may heavily depend on what you used as initial certainties, or *prior probabilities*. And, if those initial certainties are not those of a colleague, that colleague may very well assign a different certainty to the conclusion than you derived.

Because the posterior probabilities emanating from a Bayesian inference depend on the person supplying the initial certainties, and so may vary across individuals, the inferences are said to be *subjective*. This subjectivity of Bayesian inference is often mistaken for a subjective treatment of truth. Not only is such a view of Bayesianism incorrect, but it is diametrically opposed to Bayesian philosophy. The Bayesian approach represents a constructive attempt to deal with the dilemma that scientific laws and facts should not be treated as known with certainty, whereas classical deductive logic yields conclusions only when some law, fact, or connection is asserted with 100% certainty.

A common criticism of Bayesian philosophy is that it diverts attention away from the classical goals of science, such as the discovery of how the world works, toward psychological states of mind called "certainties," "subjective probabilities," or "degrees of belief" (Popper 1959). This criticism fails, however, to recognize the importance of a scientist's state of mind in determining what theories to test and what tests to apply.

Most epidemiologists desire some interval estimate or evaluation of the likely range for an effect in light of available data. This estimate must inevitably be derived in the face of considerable uncertainty about methodological details and various events that led to the available data and can be extremely sensitive to the reasoning used in its derivation. Psychological investigations have found that most people, including scientists, reason poorly in general and especially poorly in the face of uncertainty (Kahnemann et al. 1982; Gilovich et al. 2002). Bayesian philosophy provides a methodology for such reasoning and in particular provides many warnings against being overly certain about one's conclusions:

> Such warnings are echoed in refutationist philosophy. As Peter Medawar put it,
> I cannot give any scientist of any age better advice than this: the intensity of the conviction that a hypothesis is true has no bearing on whether it is true or not (Medawar 1979).

We would only add that intensity of conviction that a hypothesis is false has no bearing on whether it is false or not.

3.2.9 Impossibility of Proof

Vigorous debate is a characteristic of modern scientific philosophy, no less in epidemiology than in other areas (Rothman 1988). Perhaps the most important common thread that emerges from the debated philosophies is Hume's legacy that proof is impossible in empirical science. This simple fact is especially important to epidemiologists, who often face the criticism that proof is impossible in epidemiology, with the implication that it is possible in other scientific disciplines. Such criticism may stem from a view that experiments are the definitive source of scientific knowledge. Such a view is mistaken on at least two counts. First, the non-experimental nature of a science does not preclude impressive scientific discoveries; the myriad examples include plate tectonics, the evolution of species, planets orbiting other stars, and the effects of cigarette smoking on human health. Even when they are possible, experiments (including randomized trials) do not provide anything approaching proof and in fact may be controversial, contradictory, or irreproducible. The cold-fusion debacle demonstrates well that neither physical nor experimental science is immune to such problems (Taubes 1993).

Some experimental scientists hold that epidemiological relations are only suggestive and believe that detailed laboratory study of mechanisms within single individuals can reveal cause-effect relations with certainty. This view overlooks the fact that *all* relations are suggestive in exactly the manner discussed by Hume: Even the most careful and detailed mechanistic dissection of individual events cannot provide more than associations, albeit at a finer level. Laboratory studies often involve a degree of observer control that cannot be approached in epidemiology; it is only this control, not the level of observation, that can strengthen the inferences from laboratory studies. And again, such control is no guarantee against error.

All of the fruits of scientific work, in epidemiology or other disciplines, are at best only tentative formulations of a description of nature, even when the work itself is carried out without mistakes. The tentativeness of our knowledge does not prevent practical applications, but it should keep us skeptical and critical, not only of everyone else's work, but of our own as well. Avoidance of dogmatism or absolute certainty, along with judicious application of common sense, is arguably the most important and least controversial hallmark of the scientific enterprise (Haack 2003).

3.2.10 Causal Inference in Epidemiology

Biological knowledge about epidemiological hypotheses is often scant, making the hypotheses themselves at times little more than vague statements of causal association between exposure and disease, such as "smoking causes cardiovascular disease." These vague hypotheses have only vague consequences that can be difficult to test. To cope with this vagueness, epidemiologists usually focus on testing the negation of the causal hypothesis, that is, the null hypothesis that the exposure does *not* have a causal relation to disease. Then, any observed association can potentially refute the hypothesis, subject to the assumption (auxiliary hypothesis) that biases are absent.

If the causal mechanism is stated specifically enough, epidemiological observations can provide crucial tests of competing non-null causal hypotheses. For example, when toxic shock syndrome was first studied, there were two competing hypotheses about the origin of the toxin. Under one hypothesis, the toxin was a chemical in the tampon, so that women using tampons were exposed to the toxin directly from the tampon. Under the other hypothesis, the tampon acted as a culture medium for Staphylococci that produced the toxin. Both hypotheses explained the relation of toxic shock occurrence to tampon use. The two hypotheses, however, lead to opposite predictions about the relation between the frequency of changing tampons and the risk of toxic shock. Under the hypothesis of a chemical intoxication, more frequent changing of the tampon would lead to more exposure to the toxin and possible absorption of a greater overall dose. This hypothesis predicted that women who changed tampons more frequently would have a higher risk than women who changed tampons infrequently. The culture-medium hypothesis predicts that the women who change tampons frequently would have a lower risk than those who leave the tampon in for longer periods, because a short duration of use for each tampon would prevent the staphylococci from multiplying enough to produce a damaging dose of toxin. Thus, epidemiological research, which showed that infrequent changing of tampons was associated with the risk of toxic shock, refuted the chemical theory.

Another example of a theory easily tested by epidemiological data related to the finding that women who took replacement estrogen therapy were at a considerably higher risk for endometrial cancer. Horwitz and Feinstein (1978) conjectured a competing theory to explain the association: They proposed that women taking estrogen experienced symptoms such as bleeding that induced them to consult a physician. The resulting diagnostic workup led to the detection of endometrial cancer at an earlier stage in these women, as compared with women not taking estrogens. Many epidemiological observations could have been and were used to evaluate these competing hypotheses. The causal theory predicted that the risk of endometrial cancer would tend to increase with increasing use (dose, frequency, and duration) of estrogens, as for other carcinogenic exposures. The detection bias theory, on the other hand, predicted that women who had used estrogens only for a short while would have the greatest risk, since the symptoms related to estrogen use that led to the medical consultation tend to appear soon after use begins. Because the association of recent estrogen use and endometrial cancer was the same in both long-term and short-term estrogen users, the detection bias theory was refuted as an explanation for all but a small fraction of endometrial cancer cases occurring after estrogen use. (Refutation of the detection bias theory also depended on many other observations. Especially important was the theory's implication that there must be a large reservoir of undetected endometrial cancer in the typical population of women to account for the much greater rate observed in estrogen users.)

The endometrial cancer example illustrates a critical point in understanding the process of causal inference in epidemiological studies: Many of the hypotheses being evaluated in the interpretation of epidemiological studies are non-causal hypotheses, in the sense of involving no causal connection between the study exposure and the disease. For example, hypotheses that amount to explanations of how specific types of bias could have led to an association between exposure and disease are the usual alternatives to the primary study hypothesis that the epidemiologist needs to consider in drawing inferences. Much of the interpretation of epidemiological studies amounts to the testing of such non-causal explanations for observed associations.

3.2.11 Causal Criteria

In practice, how do epidemiologists separate out the causal from the non-causal explanations? Despite philosophic criticisms of inductive inference, inductively-oriented causal criteria have commonly been used to make such inferences. If a set of necessary and sufficient causal criteria could be used to distinguish causal from non-causal relations in epidemiological studies, the job of the scientist would be eased considerably. With such criteria, all the concerns about the logic or lack thereof in causal inference could be forgotten: It would only be necessary to consult the checklist of criteria to see if a relation were causal. We know from philosophy that a set of sufficient criteria does not exist. Nevertheless, lists of causal criteria have become popular, possibly because they seem to provide a road map through complicated territory.

Hill (1965) proposed a famous set of considerations for use in making causal inferences, which he took care not to label as "criteria." Hill's set can be seen as expanding on considerations in the landmark Surgeon General's report on Smoking and Health (1964), which in turn were anticipated by the inductive canons of Mill (1843) and discussions given by Hume. Hill suggested that the following aspects of an association be considered in attempting to distinguish causal from non-causal associations: (1) strength, (2) consistency, (3) specificity, (4) temporality, (5) biological gradient, (6) plausibility, (7) coherence, (8) experimental evidence, and (9) analogy.

Hill recognized that (apart from temporality) there is no necessary or sufficient criterion for determining whether an observed association is causal and emphasized due caution in application of his considerations (Phillips and Goodman 2004). His caution accords with the views of Hume, Popper, and others that causal inferences cannot attain the certainty of logical deductions. Although some epidemiologists have argued that it is detrimental to cloud the inferential process by considering checklist criteria (Lanes and Poole 1984), a moderate approach seeks to transform informal considerations like Hill's into deductive tests of causal hypotheses (Maclure 1985; Weed 1986; Susser 1991; Weiss 2002). Such an approach avoids the temptation to use causal criteria simply to buttress pet theories at hand and instead allows epidemiologists to focus on evaluating competing causal theories using crucial observations.

3.3 Measures of Disease Frequency

A central task in epidemiological research is to quantify the occurrence of disease in populations. We discuss here four basic measures of disease occurrence. *Incidence times* are simply the times at which new disease occurs among population members. *Incidence rate* measures the occurrence of new disease per unit of person-time. *Incidence proportion* measures the proportion of people who develop new disease during a specified period of time. *Prevalence*, a measure of status rather than of newly occurring disease, measures the proportion of people who have disease at a specific time.

3.3.1 Incidence Time

In attempting to measure the frequency of disease occurrence in a population, it is insufficient merely to record the number of people or even the proportion of the population that is affected. It is also necessary to take into account the time elapsed before disease occurs as well as the period of time during which events are counted. Consider the frequency of death. Since all people are eventually affected, the time from birth to death becomes the determining factor in the rate of occurrence of death. If, on average, death comes earlier to the members of one population than to members of another population, it is natural to say that the first population has a higher death rate than the second.

In an epidemiological study, we may measure the time of events in an individual's life relative to any one of several reference events. Using age, for example, the reference event is birth, but we might instead use the start of a treatment or the start of an exposure as the reference event. The reference event may be unique to each person, as it is with birth, or it may be identical for all persons, as with calendar time. The time of the reference event determines the time origin or *zero time* for measuring time of events.

Given an outcome event or "incident" of interest, a person's *incidence time* for this outcome is defined as the time span from zero time to the time at which the event occurs if it occurs. A man who experienced his first myocardial infarction (MI) in 1990 at age 50 has an incidence time of 1,990 in (Western) calendar time and an incidence time of 50 in age time. A person's incidence time is undefined if that person never experiences the event. (There is a useful convention that classifies such a person as having an unspecified incidence time that is known to exceed the last time the person could have experienced the event. Under this convention, a woman who had a hysterectomy in 1990 without ever having had endometrial cancer is classified as having an endometrial cancer incidence time greater than 1990.)

3.3.2 Incidence Rate

Epidemiologists often study events that are not inevitable or that may not occur during the period of observation. In such situations, the set of incidence times for a specific event in a population will not all be defined or observed, and another incidence measure must be sought. Ideally, such a measure would take into account the number of individuals in a population that become ill as well as the length of time contributed by all persons during the period they were in the population and events are counted.

Consider any population and a risk period over which we want to measure incidence in this population. Every member of the population experiences a specific amount of time in the population over the risk period; the sum of these times over all population members is called the total *person-time* at risk over the period. Person-time should be distinguished from clock time in that it is a summation of time that occurs simultaneously for many people, whereas clock time is not. Person-time represents the observational experience in which disease onsets can be observed. The number of new cases of disease divided by the person-time is the *incidence rate* of the population over the period:

$$\text{Incidence rate} = \frac{\text{no. disease onsets}}{\sum\limits_{\text{persons}} \text{time at risk for getting disease}}.$$

When the risk period is of fixed length Δt, the total person-time at risk over the period is equal to the average size of the population over the period, $\overline{N}$, times the length of the period, Δt. If we denote the incident number by A, it follows that the person-time rate equals $A/(\overline{N} \times \Delta t)$. This formulation makes clear that the incidence rate has units of inverse time (per year, per month, per day, etc.). The units attached to an incidence rate can be written as year^{-1}, month^{-1}, or day^{-1}.

It is an important principle that the only events eligible to be counted in the numerator of an incidence rate are those that occur to persons who are contributing time to the denominator of the incidence rate at the time that the disease onset occurs. Likewise, only time contributed by persons eligible to be counted in the numerator if they suffer an event should be counted in the denominator. The time contributed by each person to the denominator is sometimes known as the "time at risk," that is, time at risk of an event occurring. Analogously, the people who contribute time to the denominator of an incidence rate are referred to as the "population at risk."

Incidence rates often include only the first occurrence of disease onset as an eligible event for the numerator of the rate. For the many diseases that are irreversible states, such as diabetes, multiple sclerosis, cirrhosis, or death, there is at most only one onset that a person can experience. For some diseases that do recur, such as rhinitis, we may simply wish to measure the incidence of “first” occurrence, even though the disease can occur repeatedly. For other diseases, such as cancer or heart disease, the first occurrence is often of greater interest for study than subsequent occurrences in the same individual. Therefore, it is typical that the events in the numerator of an incidence rate correspond to the first occurrence of a particular disease, even in those instances in which it is possible for an individual to have more than one occurrence.

When the events tallied in the numerator of an incidence rate are first occurrences of disease, then the time contributed by each individual who develops the disease should terminate with the onset of disease. The reason is that the individual is no longer eligible to experience the event (the first occurrence can only occur once per individual), so there is no more information to obtain from continued observation of that individual. Thus, each individual who experiences the event should contribute time to the denominator up until the occurrence of the event but not afterward. Furthermore, for the study of first occurrences, the number of disease onsets in the numerator of the incidence rate is also a count of people experiencing the event, since only one event can occur per person.

An epidemiologist who wishes to study both first and subsequent occurrences of disease may decide not to distinguish between first and later occurrences and simply count all the events that occur among the population under observation. If so, then the time accumulated in the denominator of the rate would not cease with the occurrence of the event, since an additional event might occur in the same individual. Usually, however, there is enough of a biological distinction between first and subsequent occurrences to warrant measuring them separately. One approach is to define the “population at risk” differently for each occurrence of the event: The population at risk for the first event would consist of individuals who have not experienced the disease before; the population at risk for the second event, or first recurrence, would be limited to those who have experienced the event once and only once, etc. A given individual should contribute time to the denominator of the incidence rate for first events only until the time that the disease first occurs. At that point, the individual should cease contributing time to the denominator of that rate and should now begin to contribute time to the denominator of the rate measuring the second occurrence. If and when there is a second event, the individual should stop contributing time to the rate measuring the second occurrence and begin to contribute to the denominator of the rate measuring the third occurrence and so forth. In effect, measuring the rate for the sequence of events amounts to conducting a separate study for first events, second events, third events, and so forth, each with its own population at risk and distinct events.

3.3.3 Closed and Open Populations

Conceptually, we can imagine the person-time experience of two distinct types of populations, the *closed population* and the *open population*. A closed population adds no new members over time, and loses members only to death, whereas an

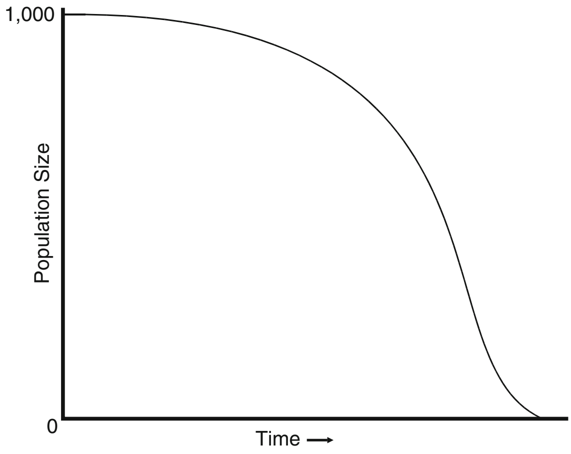


Fig. 3.3 Size of a closed population of 1,000 people, by time

open population may gain members over time through immigration or birth, or lose members who are still alive through emigration. (Some demographers and ecologists use a broader definition of closed population in which births (but not immigration or emigration) are allowed.) Suppose we graph the survival experience of a closed population of 1,000 people. Since death eventually claims everyone, after a period of sufficient time, the original 1,000 will have dwindled to zero. A graph of the size of the population with time might approximate that in Fig. 3.3.

The curve slopes downward because as the 1,000 individuals in the population die, the population at risk of death is reduced. The population is closed in the sense that we consider the fate of only the 1,000 individuals present at time zero. The person-time experience of these 1,000 individuals is represented by the area under the curve in the diagram. As each individual dies, the curve notches downward; that individual no longer contributes to the person-time denominator of the death (mortality) rate. Each individual's contribution is exactly equal to the length of time that individual is followed from start to finish; in this example, since the entire population is followed until death, the finish is the individual's death. In other instances, the contribution to the person-time experience would continue until either the onset of disease or some arbitrary cut-off time for observation, whichever came sooner.

Suppose we added up the total person-time experience of this closed population of 1,000 and obtained a total of 75,000 person-years. The death rate would be $(1{,}000/75{,}000)\times\text{year}^{-1}$ since the 75,000 person-years represent the experience of all 1,000 people until their deaths. Furthermore, if time is measured from start of follow-up, the average death time in this closed population would be 75,000 person-years/1,000 persons = 75 years, which is the inverse of the death rate.

A closed population facing a constant death rate would decline in size exponentially (which is what is meant by the term "exponential decay"). In practice, however, death rates for a closed population change with time, since the population is aging as time progresses. Consequently, the decay curve of a closed human population is never exponential. *Life-table* methodology is a procedure by which the death rate (or disease rate) of a closed population is evaluated within successive small age or time intervals, so that the age or time dependence of mortality can be elucidated. Even with such methods, it can be difficult to distinguish any age-related effects from those related to other time axes, since each individual's age increases directly with an increase along any other time axis. For example, a person's age increases with increasing duration of employment, increasing calendar time, and increasing time from start of follow-up.

An open population differs from a closed population in that individual contributions need not begin at the same time. Instead, the population at risk is open to new members who become eligible with passing time. People can enter a population open in calendar time through various mechanisms. Some are born into it; others migrate into it. For a population of people of a specific age, individuals can become eligible to enter the population by aging into it. Similarly, individuals can exit from the person-time observational experience defining a given incidence rate by dying, aging out of a defined age group, emigrating, or by becoming diseased (the latter method of exiting applies only if first bouts of a disease are being studied).

3.3.4 Steady State

If the number of people entering a population is balanced by the number exiting the population in any period of time within levels of age, sex, and other determinants of risk, the population is said to be *stationary*, or in a *steady state.* Steady state is a property that can occur only in open populations, not closed populations. It is, however, possible to have a population in steady state in which no immigration or emigration is occurring; this situation would occur if births perfectly balanced deaths in the population. The graph of the size of an open population in steady state is simply a horizontal line. People are continually entering and leaving the population in a way that might be diagrammed as shown in Fig. 3.4.

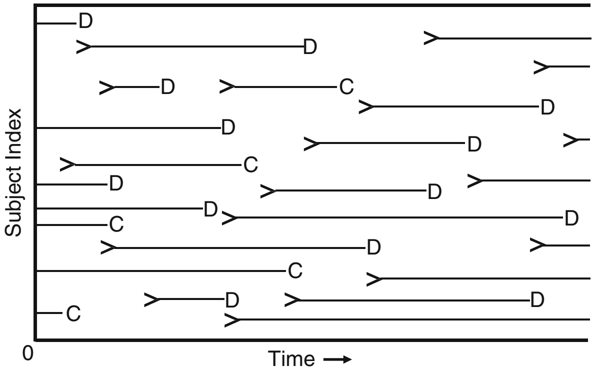


Fig. 3.4 Composition of an open population in approximate steady state, by time. > indicates entry into the population, D indicates disease onset, and C indicates exit from the population without disease

In the diagram, the symbol > represents a person entering the population, a line segment represents their person-time experience, and the termination of a line segment represents the end of their experience. A terminal D indicates that the experience ended because of disease onset, and a terminal C indicates that the experience ended for other reasons. In theory, any time interval will provide a good estimate of the incidence rate in a stationary population. The value of incidence will be the ratio of the number of cases of disease onset, indicated by D, to the area depicting the product of population × time. Because this ratio is equivalent to the density of disease onsets in the observational area, the incidence rate has also been referred to as *incidence density* (Miettinen 1976). The measure has also been called the *person-time rate, force of morbidity* (or *force of mortality* in reference to deaths), *hazard rate*, and *disease intensity*, although the latter three terms are more commonly used to refer to the theoretical limit approached by an incidence rate as the time interval is narrowed toward zero.

3.3.5 Interpretation of an Incidence Rate

The numerical portion of an incidence rate has a lower bound of zero but has no upper bound; it has the mathematical range for the ratio of two non-negative quantities, in this case the number of events in the numerator and the person-time in the denominator. At first, it may seem surprising that an incidence rate can exceed the value of 1.0, which would seem to indicate that more than 100% of a population is affected. It is true that at most, only 100% of persons in a population can get a disease, but the incidence rate does not measure the proportion of a population with illness and in fact is not a proportion at all. Recall that incidence rate is measured in units of the reciprocal of time. Among 100 people, no more than 100 deaths can occur, but those 100 deaths can occur in 10,000 person-years, in 1,000 person-years, in 100 person-years, or even in 1 person-year (if the 100 deaths occur after an average of 3.65 days each). An incidence rate of 100 cases (or deaths) per 1 person-year might be expressed as

$$100\frac{\text{cases}}{\text{person-year}}.$$

It might also be expressed as

$$10,000\frac{\text{cases}}{\text{person-century}}\text{ or}$$

$$8.33\frac{\text{cases}}{\text{person-month}}\text{ or}$$

$$1.92\frac{\text{cases}}{\text{person-week}}\text{ or}$$

$$0.27 \frac{\text{cases}}{\text{person-day}}.$$

The numerical value of an incidence rate in itself has no interpretability because it depends on the arbitrary selection of the time unit. It is thus essential in presenting incidence rates to give the appropriate time units, either as in the examples given above or as in 8.33 month^{-1} or 1.92 week^{-1}. Although the measure of time in the denominator of an incidence rate is often taken in terms of years, one can have units of years in the denominator regardless of whether the observations were collected over 1 year of time, over 2 week of time, or over 10 years of time.

The reciprocal of time is an awkward concept that does not provide an intuitive grasp of an incidence rate. The measure does, however, have a close connection to more interpretable measures of occurrence in closed populations. Referring back to Fig. 3.3, one can see that the area under the curve is equal to $N \times T$, where N is the number of people starting out in the closed population and T is the average time until death. Equivalently, the area under the curve in Fig. 3.3 is equal to the area of a rectangle with height N and width T. Since T is the average time until death for the N people, the total person-time experience is $N \times T$. The time-averaged death rate when the follow-up for the closed population is complete is $N/(N \times T) = 1/T$; that is, the death rate equals the reciprocal of the average time until death.

More generally, in a stationary population with no migration, the crude incidence rate of an inevitable outcome such as death will equal the reciprocal of the average time until the outcome. The time until the outcome is sometimes referred to as the "waiting time" until the event occurs (Morrison 1979). Thus, in a stationary population with no migration, a death rate of 0.04 year^{-1} would translate to an average time until death of 25 years.

If the outcome of interest is not death but either disease onset or death from a specific cause, the waiting time interpretation must be modified slightly: The waiting time is the average time until disease onset, assuming that a person is not at risk of other causes of death, or other events that remove one from risk of the outcome of interest. That is, the waiting time must be redefined to account for *competing risks*, that is, events that "compete" with the outcome of interest to remove persons from the population at risk.

Unfortunately, the interpretation of incidence rates as the inverse of the average "waiting time" will usually not be valid unless the incidence rate is calculated for a stationary population with no migration (no immigration or emigration) or a closed population with complete follow-up. For example, the death rate for the United States in 1977 was 0.0088 year^{-1}; in a steady state, this rate would correspond to a mean life span, or expectation of life, of 114 years. Other analyses, however, indicate that the actual expectation of life in 1977 was 73 years (Alho 1992). The discrepancy is due to immigration and to the lack of a steady state. Note that the no-migration assumption cannot hold within specific age groups, for people are always "migrating" in and out of age groups as they age.

While the notion of incidence is a central one in epidemiology, it cannot capture all aspects of disease occurrence. Consider that a rate of 1 case/(100 years) $= 0.01\ \text{year}^{-1}$

could be obtained by following 100 people for an average of 1 year and observing 1 case but could also be obtained by following two people for 50 years and observing 1 case, a very different scenario. To distinguish these situations, concepts that directly incorporate the notion of follow-up time and risk are needed.

3.3.6 Incidence Proportions and Survival Proportions

For a given interval of time, we can divide the number of new cases of disease occurring during that interval by the population size. If we measure the population size at the start of the interval and no one enters the population (immigrates) or leaves alive (emigrates) after the start of the interval, such a measure becomes the proportion of people who become cases during the time interval among those who were in the population at the start of the interval. We call this quantity the incidence *proportion*, which may also be defined as the proportion of a closed population at risk that becomes diseased within a given period of time. This quantity is often called the "cumulative incidence" (Miettinen 1976), but the term "cumulative incidence" is also used for another quantity we will discuss below. A more traditional term for incidence proportion is "attack rate," but we reserve the term "rate" for person-time incidence rates.

If *risk* is defined as the probability of an individual developing disease in a specified time interval, then incidence proportion is a measure, or estimate, of average risk. Although this concept of risk applies only to individuals and incidence proportion to populations, incidence proportion is sometimes called "risk". "Average risk" is a more accurate synonym, one that we will sometimes use.

Like any proportion, the value of an incidence proportion ranges from zero to one and is dimensionless. It is uninterpretable, however, without specification of the time period to which it applies. An incidence proportion of death of 3% means something very different when it refers to a 40-year period than when it refers to a 40-day period.

A useful complementary measure to the incidence proportion is the *survival proportion*, which may be defined as the proportion of a closed population at risk that does *not* become diseased within a given period of time. If R and S denote the incidence and survival proportions, we have that $S = 1 - R$ and $R = 1 - S$. Another measure that is commonly used is the *incidence odds*, defined as $R/S = R/(1-R)$, the ratio of the proportion getting the disease to the proportion not getting the disease. If R is small, $S \doteq 1$ and $R/S \doteq R$; that is, the incidence odds will approximate the incidence proportion when both quantities are small. Otherwise, because $S < 1$, the incidence odds will be greater than the incidence proportion.

Under certain conditions, there is a very simple relation between the incidence proportion and the incidence rate of a non-recurrent event. Consider a closed population over an interval t_0 to t_1 and let $\Delta t = t_1 - t_0$ be the length of the interval. If N is the size of the population at t_0, and A is the number of disease onsets over the interval, then the incidence and survival proportions over the interval are $R = A/N$ and $S = (N - A)/N$. Now suppose the size of the population at risk declines only slightly over the interval. Then $N - A \doteq N$, $S \doteq 1$, and so $R/S \doteq R$. Furthermore, the average size of the population at risk will be approximately N, and so the total person-time at risk over the interval will be approximately $N\Delta t$. Thus, the incidence rate (I) over the interval will be approximately $A/N\Delta t$, and we obtain

$$R = A/N = (A/N\Delta t)\Delta t \doteq I\Delta t \doteq R/S.$$

In words, the incidence proportion, incidence odds, and the quantity $I\Delta t$ will all approximate one another if the population at risk declines only slightly over the interval. We can make this approximation hold to within an accuracy of $1/N$ by making Δt so short that no more than one person leaves the population at risk over the interval. Thus, given a sufficiently short time interval, one can simply multiply the incidence rate by the time period to approximate the incidence proportion. This approximation offers another interpretation for the incidence rate: It can be viewed as the limiting value of the ratio of the average risk to the time period for the risk as the duration of the time period approaches zero.

A specific type of incidence proportion is the *case fatality rate*, or *case fatality ratio*, which is the incidence proportion of death among those who develop an illness (it is therefore not a rate in our sense but a proportion). The time period for measuring the case fatality rate is often unstated, but it is always better to specify it.

3.3.7 Prevalence

Unlike incidence measures, which focus on events, *prevalence* focuses on disease status. Prevalence may be defined as the proportion of a population that has disease at a specific point in time. The terms *point prevalence*, *prevalence proportion*, and *prevalence rate* are sometimes used to mean the same thing. The *prevalence pool* is the subset of the population with the disease. An individual who dies with or from disease is removed from the prevalence pool; consequently, death from an illness decreases prevalence. Diseases with large incidence rates may have low prevalences if they are rapidly fatal. People may also exit the prevalence pool by recovering from disease or emigrating from the population.

Recall that a stationary population has an equal number of people entering and exiting during any unit of time. Suppose that both the population at risk and the prevalence pool are stationary and that everyone is either at risk or has the disease. Then the number of people entering the prevalence pool in any time period will be balanced by the number exiting from it:

$$\text{Inflow (to prevalence pool)} = \text{outflow (from prevalence pool).}$$

People can enter the prevalence pool from the non-diseased population and by immigration from another population. Suppose there is no immigration into or emigration from the prevalence pool, so that no one enters or leaves the pool except by disease onset, death, or recovery. If the size of the population is N and the size of the prevalence pool is P, then the size of the population at risk that "feeds" the prevalence pool will be $N - P$. Also, during any time interval of length Δt, the number of people who enter the prevalence pool will be

$$I(N - P)\Delta t,$$

where I is the incidence rate, and the outflow from the prevalence pool will be

$$I' P \Delta t,$$

where I' represents the incidence rate of exiting from the prevalence pool, that is, the number who exit divided by the person-time experience of those in the prevalence pool.

3.3.8 Prevalence, Incidence, and Mean Duration

Earlier, we mentioned that, in the absence of migration, the reciprocal of an incidence rate in a stationary population equals the mean time spent in the population before the incident event. Therefore, in the absence of migration and in a stationary population, the reciprocal of I' will be the mean duration of the disease, $\overline{D}$, which is the mean time until death or recovery. It follows that

$$\text{inflow} = I(N - P)\Delta t = \text{outflow} = (1/\overline{D})P\Delta t$$

which yields

$$\frac{P}{N - P} = I \times \overline{D}.$$

$P/(N - P)$ is the ratio of diseased to non-diseased people in the population or equivalently, the ratio of the prevalence proportion to its complement (1 − prevalence proportion). (We could call those who are non-diseased healthy except that we mean they do not have a specific illness, which doesn't imply an absence of all illness.) The ratio $P/(N - P)$ is called the *prevalence odds*; it is the odds of having a disease relative to not having the disease. As shown above, the prevalence odds equals the incidence rate times the mean duration of illness. If the prevalence is small, say less than 0.1, then

$$\text{Prevalence proportion} \doteq I \times \overline{D}$$

since the prevalence proportion will approximate the prevalence odds for small values of prevalence. More generally (Freeman and Hutichison 1980), under the assumption of stationarity and no migration in or out of the prevalence pool,

$$\text{Prevalence proportion} = \frac{I \times \overline{D}}{1 + I \times \overline{D}}$$

which can be obtained from the above expression for the prevalence odds, $P/(N - P)$.

Like the incidence proportion, the prevalence proportion is dimensionless, with a range of zero to one. The above equations are in accord with these requirements, because in each of them, the incidence rate, with a dimensionality of the reciprocal of time, is multiplied by the mean duration of illness, which has the dimensionality of time, giving a dimensionless product.

Furthermore, the product $I \times \overline{D}$ has the range of zero to infinity, which corresponds to the range of prevalence odds, whereas the expression

$$\frac{I \times \overline{D}}{1 + I \times \overline{D}}$$

is always in the range zero to one, corresponding to the range of a proportion.

Unfortunately, the above formulas have limited practical utility because of the no-migration assumption and because they do not apply to age-specific prevalence (Miettinen 1976). If we consider the prevalence pool of, say, diabetics age 60–64, we can see that this pool experiences considerable immigration from younger diabetics aging into the pool and considerable emigration from members aging out of the pool. Under such conditions, we require more elaborate formulas that give prevalence as a function of age-specific incidence, duration, and other population parameters (Preston 1987; Keiding 1991; Alho 1992).

3.3.9 Utility of Prevalence in Etiological Research

Seldom is prevalence of direct interest in etiological applications of epidemiological research. Since prevalence reflects both the incidence rate and the probability of surviving with disease, studies of prevalence, or studies based on prevalent cases, yield associations that reflect the determinants of survival with disease just as much as the causes of disease. The study of prevalence can be misleading in the paradoxical situation in which better survival from a disease and therefore a higher prevalence follows from the action of preventive agents that mitigate the disease once it occurs. In such a situation, the preventive agent may be positively associated with the prevalence of disease and so be misconstrued as a causative agent.

Nevertheless, for one class of diseases, namely, congenital malformations, prevalence is usually employed. The proportion of babies born with some malformation is a prevalence proportion, not an incidence rate. The incidence of malformations refers to the occurrence of the malformations among the susceptible populations of embryos. Many malformations lead to early embryonic or fetal death that is classified, if recognized, as a miscarriage rather than a birth. Thus, malformed babies at birth represent only those individuals who survived long enough with their malformations to be recorded as a birth. This is indeed a prevalence measure, the reference point in time being the moment of birth. The measure classifies the population of newborns as to their disease status, malformed or not, at the time of birth. This example illustrates that the time reference for a prevalence need not be a common point in calendar time: It can be a point on another time scale, such as an individual's life span.

It would be more useful and desirable to study the incidence than the prevalence of congenital malformations; as already noted, studying prevalence makes it impossible to distinguish the effects of agents that increase the incidence rate from the effects of agents that increase survival with the disease once the disease occurs. Unfortunately, it is seldom possible to measure the incidence rate of malformations, since the population at risk, young embryos, is difficult to ascertain, and learning the occurrence and timing of the malformations among the embryos is equally problematic. Consequently, in this area of research, incident cases are not usually studied, with most investigators settling for the theoretically less desirable but much more practical study of prevalence at birth.

Prevalence is sometimes used to measure the occurrence of non-lethal degenerative diseases with no clear moment of onset. It is also used in seroprevalence studies of the incidence of infection, especially when the infection has a long asymptomatic (silent) phase that can only be detected by serum testing (such as HIV infection). In these and other situations, prevalence is measured simply for convenience, and inferences are made about incidence by using assumptions about the duration of illness. Of course, in epidemiological applications outside of etiological research, such as planning for and managing health resources and facilities, health economics, and other public health activities, prevalence may be a more relevant measure than incidence.

Although prevalence is generally of lesser interest than incidence for etiological research, in many public health applications prevalence is of primary interest. The reason is that the disease burden in a population is often more closely related to prevalence than to incidence. For example, a factor that has no effect on asthma incidence but which prolongs the duration of asthmatic episodes would increase the disease burden, as reflected in the need for medication, the utilization of inpatient or emergency medical services, and number of outpatient visits.

3.4 Measures of Effect

Epidemiologists use the term *effect* in two senses. In one sense, any case of a given disease may be the effect of a given cause. *Effect* is used in this way to mean the endpoint of a causal mechanism, identifying the type of outcome that a cause produces. For example, we may say that HIV infection is an effect of sharing needles for drug use. This use of the term *effect* merely identifies HIV infection as one consequence of the activity of sharing needles. Other effects of the exposure, such as Hepatitis B infection, are also possible.

In a more particular and quantitative sense, an *effect* is also the amount of change in a population's disease frequency caused by a specific factor. If disease frequency is measured in terms of incidence rate or proportion, then the effect is the change in incidence rate or proportion brought about by a particular factor. We might say that for drug users, the effect of sharing needles, compared with not sharing needles, is to increase the average risk of HIV infection from 0.001 in 1 year to 0.01 in 1 year. Although it is customary to use the definite article in referring to this second type of effect ("the" effect of sharing needles), it is not meant to imply that this is a unique effect of sharing needles. An increase in risk for Hepatitis or other diseases remains possible, and the increase in risk of HIV infection may differ across populations and time.

In epidemiology, it is customary to refer to potential causal characteristics as *exposures*. Thus, "exposure" can refer to a behavior (such as needle sharing), a treatment (such as an educational program about hazards of needle sharing), a trait (such as genotype), or an exposure in the ordinary sense (such as injection of contaminated blood).

Population effects are most commonly expressed as effects on incidence rates or incidence proportions, but other measures based on the incidence times or prevalences may also be used. Epidemiological analyses that focus on survival time until death or recurrence of disease are examples of analyses that measure effects on incidence times. *Absolute effects* are differences in incidence rates, incidence proportions, prevalences, or incidence times. *Relative effects* involve ratios of these measures.

3.4.1 Simple Effect Measures

Consider a cohort followed over a specific time interval – say 1996–2000, or age 50–69. If we can imagine the experience of this cohort over the same interval under two different conditions – say, "exposed" and "unexposed" – then we can ask what the incidence rate of any outcome would be under the two conditions. Thus, we might consider a cohort of smokers and an exposure that consisted of mailing to each cohort member a brochure of current smoking cessation programs in their county of residence. We could then ask what the lung cancer incidence rate would be in this cohort if we carry out this treatment and what it would be if we did not carry out this treatment. The difference between the two rates we call the absolute effect of our mailing program on the incidence rate, or the *causal rate difference*. To be brief, we might refer to the causal rate difference as the excess rate due to the program (which would be negative if the program prevented some lung cancers).

In a parallel manner, we may ask what the incidence proportion would be if we carry out this treatment and what it would be if we do not carry out this treatment. The difference of the two proportions we call the absolute effect of our treatment on the incidence proportions, or *causal risk difference* or excess risk for short. Also in a parallel fashion, the difference in the average lung-cancer-free years of life lived over the interval under the treated and untreated conditions is another absolute effect of treatment.

To illustrate the above measures in symbolic form, suppose we have a closed cohort of size N at the start of a fixed time interval and that anyone alive without the disease is at risk of the disease. Further, suppose that if every member of the cohort gets exposed throughout the interval, A_1 cases will occur and the total time at risk will be T_1, but if no member of the same cohort is exposed during the interval, A_0 cases will occur and the total time at risk will be T_0. Then the causal rate difference will be

$$\frac{A_1}{T_1} - \frac{A_0}{T_0},$$

the causal risk difference will be

$$\frac{A_1}{N} - \frac{A_0}{N},$$

and the causal difference in average disease-free time will be

$$\frac{T_1}{N} - \frac{T_0}{N}.$$

Each of these measures compares disease occurrence by taking differences and so are called difference measures, or absolute measures.

More commonly, effect measures are defined by taking ratios. Examples of such ratio (or relative) measures are the *causal rate ratio*

$$\frac{A_1/T_1}{A_0/T_0} = \frac{I_1}{I_0},$$

where $I_j = A_j/T_j$ is the incidence rate under condition j ($1 =$ exposed, $0 =$ unexposed); the *causal risk ratio*

$$\frac{A_1/N}{A_0/N} = \frac{A_1}{A_0} = \frac{R_1}{R_0},$$

where $R_j = A_j/N$ is the incidence proportion (average risk) under condition j; and the *causal ratio of disease-free time*,

$$\frac{T_1/N}{T_0/N} = \frac{T_1}{T_0}.$$

The rate ratio and risk ratio are often called relative risk measures. The three ratio measures are related by the simple formula

$$\frac{R_1}{R_0} = \frac{R_1 N}{R_0 N} = \frac{A_1}{A_0} = \frac{I_1}{I_0}\frac{T_1}{T_0},$$

which follows from the fact that the number of cases equals the disease rate times the time at risk. A fourth relative risk measure can be constructed from the incidence odds. If we write $S_1 = 1 - R_1$ and $S_0 = 1 - R_0$, the *causal odds ratio* is then

$$\frac{R_1/S_1}{R_0/S_0} = \frac{A_1/(N - A_1)}{A_0/(N - A_0)}.$$

The definitions of effect just given are sometimes called *counterfactual* or *potential-outcome* definitions. Such definitions may be traced back to the writings of Hume but received little attention from scientists until the twentieth century (see Lewis 1973; Rubin 1990 for early references in philosophy and statistics, respectively). The definitions are often called "counterfactual" because at least one of the two circumstances in the definitions must be contrary to fact: The cohort may be exposed or "treated" (e.g., every member sent a mailing) or untreated (no one sent a mailing); if the cohort is treated, then the untreated condition will be counterfactual, and if it is untreated, then the treated condition will be counterfactual. Both conditions may be counterfactual: If only part of the cohort is sent the mailing, both conditions in the definitions will be contrary to this fact. Nonetheless, one of the two conditions compared may be the actual treatment received by the population, in which case the outcome under that treatment is not counterfactual. Hence, the term "potential-outcome" better captures the nature of the comparisons, which are between outcomes that could have occurred (depending on the treatment given), and one of which may have occurred.

Potential-outcome comparisons amount to thought experiments, useful for conceptualizing what is meant by causal effects. In practice, actual estimation of causal effects involves circumstances that only simulate counterfactual thought experiments, as explained below. Although some authors have objected to counterfactual causal concepts on philosophical grounds, it turns out that such concepts directly parallel concepts found in graphical and structural-equation models of causality (Pearl 2009; Greenland and Brumback 2002).

An important feature of potential-outcome definitions of effect is that they involve two distinct conditions: an index condition, which usually involves some exposure or treatment, and a reference condition against which this exposure of treatment will be evaluated – such as no treatment. To ask for "the" effect of exposure is meaningless without reference to some other condition. In the above example, the effect of one mailing is only defined in reference to no mailings. We could have instead asked about the effect of one mailing relative to four mailings; this is a very different comparison than one versus no mailing.

3.4.2 Effect-Measure Modification

Suppose we divide our cohort into two or more distinct categories, or strata. In each stratum, we can construct an effect measure of our choosing. These stratum-specific effect measures may or may not equal one another. Rarely would we have any reason to suppose that they do equal one another. If indeed they are not equal, we say that the effect measure is *heterogeneous*, *modified*, or *varies* across strata. If they are equal, we say that the measure is *homogeneous*, *uniform*, or *constant* across strata.

A major point about effect-measure modification is that, if effects are present, it will usually be the case that only one or none of the effect measures discussed above will be uniform across strata. In fact, if the exposure has any effect on an occurrence measure, at most one of the ratio or difference measures of effect can be uniform across strata. As an example, suppose among males the average risk would be 0.50 if exposure was present but would be 0.20 if exposure was absent, whereas among females, the average risk would be 0.10 if exposure was present but would be 0.04 if exposure was absent. Then the causal risk difference for males is $0.50-0.20 = 0.30$, five times the female difference of $0.10 - 0.04 = 0.06$. In contrast, for both sexes, the causal risk ratio is $0.50/0.20 = 0.10/0.04 = 2.5$. Now suppose we change this example to make the differences uniform, say by making the exposed male risk 0.26 instead of 0.50. Then both differences would be 0.06, but the male risk ratio would be $0.26/0.20 = 1.3$, much less than the female risk ratio of 2.5.

3.4.3 Relative Versus Absolute Measures

As mentioned above, we refer to differences in incidence rates, incidence proportions, prevalences, or incidence times as absolute measures. Relative effect measures are based on the ratio of an absolute effect measure to a baseline measure of occurrence. Analogous measures are used routinely whenever change or growth is measured. For example, suppose that an investment of a sum of money has yielded a gain of $1,000 in 1 year. Knowing the gain might be useful in itself for some purposes, but the absolute increase in value does not reveal by itself how effective the investment was. If the initial investment was $5,000 and grew to $6,000 in 1 year, then we could judge the investment by relating the absolute gain, $6,000–$5,000, to the initial amount. That is, we take the $1,000 gain and divide it by the $5,000 of the original principal, obtaining 20% as the relative gain. The relative gain puts the absolute gain into a perspective that reveals how effective the investment was.

Because the magnitude of the relative effect depends on the magnitude of the baseline occurrence, the same absolute effect in two populations can correspond to greatly differing relative effects (Peacock 1971). Conversely, the same relative effects for two populations could correspond to greatly differing absolute effects.

3.4.4 Attributable Fractions

Although the potential-outcome approach to effects has provided the foundation for extensive statistical and philosophical developments in causal analysis, it takes no account of the mechanisms that produce effects. Suppose that all sufficient causes of a particular disease were divided into two sets, those that contain a specific cause (exposure) and those that do not, and that the exposure is never preventive. This situation is summarized in Fig. 3.5.

C and C′ may represent many different combinations of causal components. Each of the two sets of sufficient causes represents a theoretically large variety of causal

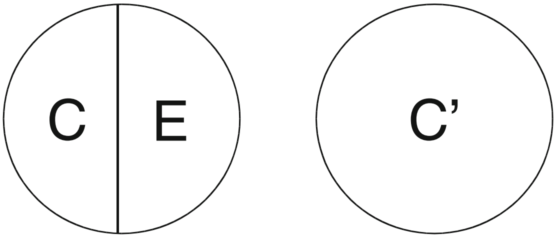


Fig. 3.5 Two types of sufficient causes of a disease

mechanisms for disease, perhaps as many as one distinct mechanism for every case that occurs. Disease can occur either with or without E, the exposure of interest. The causal mechanisms are grouped in the diagram according to whether or not they contain the exposure. We say that the exposure E causes disease if a sufficient cause that contains E gets completed. Thus, we say that exposure can cause disease if exposure will cause disease under at least some set of conditions C.

Perhaps the most straightforward way to quantify the effect of the exposure would be to estimate the numbers of cases that were caused by E. This number is not estimable from ordinary incidence data, because the observation of an exposed case does not reveal the mechanism that caused the case. In particular, people who have the exposure can develop the disease from a mechanism that does not include the exposure. For example, a smoker may develop lung cancer through some mechanism that does not involve smoking (e.g., one involving asbestos or radiation exposure). For such lung cancer cases, their smoking was incidental; it did not contribute to the cancer causation. There is currently no way to tell which exposures are responsible for a given case. Therefore, exposed cases include some cases of disease caused by the exposure, if the exposure is indeed a cause, and some cases of disease that occur through mechanisms that do not involve the exposure.

The observed incidence rate or proportion among the exposed reflects the incidence of cases in both sets of sufficient causes represented in Fig. 3.5. The incidence of sufficient causes containing E could be found by subtracting the incidence of the sufficient causes that lack E. Unfortunately, the latter incidence cannot be estimated if we cannot distinguish cases for which exposure played an etiological role from cases for which exposure was irrelevant (Greenland and Robins 1988). Thus, if I_1 is the incidence rate of disease in a population when exposure is present, and I_0 is the rate in that population when exposure is absent, the rate difference $I_1 - I_0$ does not necessarily equal the rate of disease with the exposure as a component cause.

To see the source of this difficulty, imagine a cohort in which, for every member, the causal complement of exposure, C, will be completed before the sufficient cause C′ is completed. If the cohort is unexposed, every case of disease must be attributable to the cause C′. But, if the cohort is exposed from start of follow-up, every case of disease occurs when C is completed (E being already present), so every case of disease must be attributable to the sufficient cause containing C and E. Thus, the incidence rate of cases caused by exposure is I_1 when exposure is present, not $I_1 - I_0$.

Several other measures have often been incorrectly interpreted as the fraction of cases caused by exposure, or etiological fraction. One such measure is the *rate fraction*,

$$\frac{I_1 - I_0}{I_1} = \frac{I_1/I_0 - 1}{I_1/I_0} = \frac{IR - 1}{IR},$$

also known as the *relative excess rate,* in which IR denotes the incidence rate ratio. The preceding example shows that the rate fraction is generally *not* equal to the fraction of cases in which exposure played a role in the disease etiology, for in the example, the latter fraction is 100%. Another fractional measure is $(A_1 - A_0)/A_1$, the excess caseload due to exposure, which has been called the *excess fraction* (Greenland and Robins 1988). The preceding example shows that this measure may be far less than the etiological fraction, for in the example, the latter fraction is 100%, regardless of A_0.

There has been much confusion in the epidemiological literature over the definition and interpretation of terms related to the above concepts. The term "attributable risk" has, at one time or another, been used to refer to the risk difference, the rate fraction, the etiological fraction, and the excess fraction. The terms "etiological fraction," "attributable fraction," and "attributable proportion" have each been used to refer to the etiological fraction at one time, the excess fraction at others, and the rate fraction at still other times.

In a closed cohort, the fraction of the exposed incidence proportion $R_1 = A_1/N$ that is attributable to exposure is exactly equal to the excess fraction:

$$\frac{R_1 - R_0}{R_1} = \frac{A_1/N - A_0/N}{A_1/N} = \frac{A_1 - A_0}{A_1},$$

where $R_0 = A_0/N$ is what the incidence proportion would be with no exposure. An equivalent formula for the excess fraction is

$$\frac{R_1 - R_0}{R_1} = \frac{R_1/R_0 - 1}{R_1/R_0} = \frac{RR - 1}{RR},$$

where RR is the causal risk ratio. The rate fraction is often mistakenly equated with either the etiological fraction or the excess fraction. To see that it is not equal to the excess fraction, let T_1 and T_0 represent the total time at risk that would be experienced by the cohort under exposure and non-exposure during the interval of interest. The rate fraction $(I_1 - I_0)/I_1$ then equals

$$\frac{A_1/T_1 - A_0/T_0}{A_1/T_1}.$$

If exposure has any effect, and if the disease removes people from further risk (as when the disease is irreversible), then T_1 will be less than T_0. The last expression cannot equal the excess fraction $(A_1 - A_0)/A_1$ if $T_1 \neq T_0$, although if the exposure effect on total time at risk is small, T_1 will be close to T_0 and so the rate fraction will approximate the excess fraction. Although the excess fraction and rate fraction for an uncommon disease will usually be close to one another, for reasons outlined above, both may be much less than the etiological fraction (Greenland and Robins 1988). This discrepancy leads to some serious issues in policy and legal settings, in which the etiological fraction corresponds to the probability of causation (PC), that is, the probability that the disease of a randomly selected case had exposure as a component cause. In these settings, an estimate of the excess fraction or rate fraction in the exposed is often presented as an estimate of PC. Unfortunately, the excess fraction and rate fraction can be considerably different from the PC, even if the disease is rare; hence, such presentations are misleading (Greenland 1999; Greenland and Robins 2000).

For convenience, we refer to the family of fractional measures, including the etiological, excess, and rate fractions as "attributable fractions" as in Rothman et al. (2008). This term was originally introduced by Ouellet et al. (1979) and Deubner et al. (1980). These fractions are also often called "attributable risk percent" (Cole and MacMahon 1971; Koepsell and Weiss 2003) or "attributable risk," although the latter term is also used to denote the risk difference (MacMahon and Pugh 1970; Koepsell and Weiss 2003). These measures were intended for use with exposures that have a net causal effect; they become negative and hence difficult to interpret with a net preventive effect. One simple approach for dealing with preventive exposures is to interchange the exposed and unexposed quantities in the above formulas, interchanging I_1 with I_0, P_1 with P_0, A_1 with A_0, and T_1 with T_0. The resulting measures have been called *preventable fractions* and are easily interpreted. For example, $(A_0 - A_1)/A_0 = (R_0 - R_1)/R_0 = 1 - R_1/R_0 = 1 - RR$ is the fraction of the caseload under non-exposure that could be prevented by exposure.

3.4.5 Population Attributable Fractions and Impact Fractions

One often sees in the literature a definition of "population attributable risk" or "population attributable fraction" as the reduction in incidence that would be achieved if the population had been entirely unexposed, compared with its current (actual) exposure pattern. One should recognize that this concept, due to Levin (1953), is just a special case of the definition of attributable fraction based on exposure pattern. In particular, it is a comparison of the incidence (either rate or number of cases, which must be kept distinct) under the observed pattern of exposure, and the incidence under a counterfactual pattern in which exposure or treatment is entirely absent from the population. A more general concept is the "impact fraction" (Morgenstern and Bursic 1982), which is a comparison of incidence under the observed exposure pattern and incidence under a counterfactual pattern in which exposure is only partially removed from the population. Again, this is a special case of our definition of attributable fraction based on exposure pattern.

3.4.6 Estimation of Effects

Effects are defined in reference to a *single*, enumerable population, under two distinct conditions. Such definitions require that one can meaningfully describe each condition for the one population. Consider, for example, the "effect" of sex (male versus female) on heart disease. For these words to have content, we must be able to imagine a cohort of men, their heart disease incidence, and what their incidence would have been had the very same men been women instead. The apparent ludicrousness of this demand reveals the vague meaning of sex effect. To reach a reasonable level of scientific precision, sex effect could be replaced by more precise mechanistic concepts, such as hormonal effects and effects of other sex-associated factors. With such concepts, we can imagine what it means for the men to have their exposure changed: hormone treatments, sex-change operations, and so on.

The single population in an effect definition can only be observed under one of the two conditions in the definition (and sometimes neither). This leads to the problem of effect *estimation*, which is to predict accurately what the magnitude of disease occurrence would have been in the single population under conditions that did not in fact occur (counterfactual conditions). For example, we may have observed $I_1 = 50$ deaths/100,000 person-years in a target cohort of smokers over a 10-year follow-up and ask what rate reduction would have been achieved had these smokers quit at the start of follow-up. Here, we observed a rate I_1 and are asking about I_0, the rate that would have occurred under complete smoking cessation.

Since I_0 is not observed, we must predict what it would have been. To do so, we would want to refer to outside data such as data from a cohort that was not part of the target cohort. From these data, we would construct a prediction of I_0. The point we wish to emphasize here is that neither the outside cohort nor the prediction derived from it are part of the effect measure: They are only ingredients in our estimation process. This point is overlooked by effect definitions that refer to two separate "exposed" and "unexposed" populations. Such definitions confuse the concept of effect with the concept of *association*.

3.4.7 Measures of Association

A measure of association, as opposed to a measure of effect, is a contrast between occurrence measures in two different populations or in a single population at two different times. For example, if we took the ratio of cancer incidence rates among males and females in Canada, it would not be an effect measure, because its two component rates refer to different groups of people. We describe the rate ratio as a *measure of association*; in this example, it is a measure of the association of gender with cancer incidence in Canada. As another example, we could contrast the incidence rate of dental caries in a community in the year before and in the third year after the introduction of fluoridation of the water supply. If we take the difference of the rates in these before and after periods, this difference is not an effect measure, because its two component rates refer to two different populations, one before fluoridation and one after. There may be considerable or even complete overlap in the persons present in the before and after periods; nonetheless, the experiences compared refer to different time periods. The rate difference is a measure of the association of fluoridation with dental caries incidence in the community, but it is not an effect measure.

All observable epidemiological contrasts between occurrence measures should be considered measures of association, rather than effect measures, because all epidemiological measures that contrast two or more observed occurrence measures must involve some differences in persons or time between the measures; we cannot observe a single population undergoing two mutually exclusive exposures (e.g., smoking, not smoking) at the same time. It is only as a thought experiment that we can imagine a comparison among potential outcomes, because by definition, at least one (and possibly both) occurrence measures being compared will be counterfactual and hence unobserved for the population. Nonetheless, if we can obtain estimates of occurrence measures in that population under those two events or states, either by direct observation or by a prediction from data outside the population, we can estimate effect measures.

The more difficult issue is whether the imagined effect measure is a well-defined contrast for a single person or population. To imagine an effect, we must imagine a single population that could have theoretically experienced either of two contrasting events or states (e.g., smoking and not smoking) as possible alternative histories (Greenland 2002, 2005; Hernán 2005). This challenge may be minor for medical treatments and similar interventions, for usually, it is no trouble to imagine a treated person as having been untreated instead and vice-versa. The situation is far different when discussing personal characteristics. For example, when talking of sex effects, it is at least highly ambiguous to talk of what would have happened to a woman had she been a man instead, since it leaves open too many crucial details. Does "being a man instead" mean a sex-change operation? If so, the effect under study is really that of the operation. Or does "being a man instead" mean merely using men's hairstyle and clothing well enough to be mistaken for a man? If so, the effect under study is really that of effective change in appearance.

3.5 Confounding

In the preceding example of dental caries, it is tempting to ascribe any and all of a decline in incidence following fluoridation to the act of fluoridation itself. Let us analyze what such an inference translates into in terms of measures of effect and association. The *effect* we wish to measure is that which fluoridation had on the rate; to measure this effect, we must contrast the actual rate under fluoridation with the rate that would have occurred *in the same time period* had fluoridation *not* been introduced. We cannot observe the latter rate, for it is counterfactual. Thus, we substitute in its place, or exchange, the rate in the time period before fluoridation.

In doing so, we substitute a measure of association (the rate difference before and after fluoridation) for what we are really interested in (the causal rate difference between rates without and with fluoridation in the post-fluoridation time period).

This substitution will be misleading to the extent that the rate before fluoridation does not equal and so should not be exchanged with the counterfactual rate (i.e., the rate that would have occurred in the post-fluoridation period if fluoridation had not been introduced). If the two are not equal, then the measure of association we are using will not equal the measure of effect we are substituting it for. In such a circumstance, we say that our measure of association is *confounded* (for our desired measure of effect). Other ways of expressing the same idea is that the before-after rate difference is confounded for the causal rate difference or that *confounding* is present in the before-after difference (Greenland and Robins 1986). On the other hand, if the rate before fluoridation does equal the counterfactual rate, so that the measure of association equals our desired measure of effect, we say that the before-after difference is unconfounded or that no confounding is present in this difference.

The preceding definitions apply to ratios as well as differences. Because ratios and differences contrast the same underlying quantities, confounding of a ratio measure implies confounding of the corresponding difference measure and vice versa: If the value substituted for the counterfactual rate or risk does not equal that rate or risk, both the ratio and difference will be confounded.

The above definitions also extend immediately to situations in which the contrasted quantities are average risks, incidence times, or prevalences. For example, one could wish to estimate the impact of fluoridation on caries prevalence 3 years after fluoridation began. Here, the needed but unobserved counterfactual is what the caries prevalence would have been 3 years after fluoridation began had fluoridation not begun; for it, we might substitute the prevalence of caries at the time fluoridation began. It is possible (though perhaps rare in practice) for there to be confounding for one effect measure and not another if the two effect measures derive from different underlying occurrence measures. For example, there could in theory be confounding of the rate ratio but not the risk ratio.

One point of confusion in the literature is the failure to recognize that odds are risk-based measures, and hence odds ratios will be confounded under exactly the same circumstances as risk ratios (Miettinen and Cook 1981; Greenland and Robins 1986; Greenland 1987). The confusion arises because of the peculiarity that the causal odds ratio for a whole cohort can be closer to the null than any stratum-specific causal odds ratio. Such non-collapsibility of the causal odds ratio is usually confused with confounding, even though it has nothing to do with the latter phenomenon (Greenland et al. 1999).

Consider again the fluoridation example. Suppose that, within the year after fluoridation began, dental hygiene education programs were implemented in some of the schools in the community. If these programs were effective, then (other things being equal) some reduction in caries incidence would have occurred as a consequence of the programs. Thus, even if fluoridation had not begun, the caries incidence would have declined in the post-fluoridation time period. In other words, the programs alone would have caused the counterfactual rate in our effect measure to be lower than the pre-fluoridation rate that substitutes for it. As a result, the measure of association (which is the before-after rate difference) must be larger than the desired measure of effect (the causal rate difference). In this situation, we say the programs *confounded* the measure of association or that the program effects are confounded with the fluoridation effect in the measure of association. We also say that the programs are *confounders* of the association and that the association is confounded by the programs.

Confounders are factors (exposures, interventions, treatments, etc.) that explain or produce confounding. In the present example, the programs explain why the before-after association overstates the fluoridation effect: The before-after risk difference or ratio includes the effects of programs as well as the effects of fluoridation. More generally, a confounder explains a discrepancy between the desired but unobservable counterfactual risk or rate (which the exposed would have had, had they been unexposed) and the unexposed risk or rate that was its substitute. In order for a factor to explain some of this discrepancy, and thus confound, it must be capable of affecting or at least predicting the risk or rate in the unexposed (reference) group. In the above example, we assumed that the presence of the dental hygiene programs in the years after fluoridation accounted for some of the discrepancy between the before-fluoridation rate and the (counterfactual) rate that would have occurred 3 years after fluoridation if fluoridation had not been introduced.

A large portion of epidemiological methods are concerned with avoiding or adjusting (controlling) for confounding. Such methods inevitably rely on the gathering and proper use of confounder measurements. The most fundamental adjustment methods rely on the notion of *stratification* on confounders. If we make our comparisons within specific levels of a confounder, those comparisons cannot be confounded by that confounder. For example, we could limit our before-after fluoridation comparisons to schools in states in which no dental-hygiene program was introduced. In such schools, program introductions could not have had an effect (because no program was present), and so any decline following fluoridation could not be explained by effects of programs in those schools.

3.6 Selection Bias

Selection biases are distortions that result from procedures used to select subjects and from factors that influence study participation. The common element of such biases is that the relation between exposure and disease is different for those who participate and those who should be theoretically eligible for study, including those who do not participate. The result is that associations observed in the study represent a mix of forces determining participation as well as forces determining disease.

3.6.1 Self-Selection Bias

One form of such bias is self-selection bias. When the Centers for Disease Control (CDC) investigated subsequent leukemia incidence among troops who had been present at the Smoky Atomic Test in Nevada (Caldwell et al. 1980), 76% of the troops identified as members of that cohort had known outcomes. Of this 76%, 82% were traced by the investigators, but the other 18% contacted the investigators on their own initiative in response to publicity about the investigation. This self-referral of subjects is ordinarily considered a threat to validity, since the reasons for self-referral may be associated with the outcome under study (Criqui et al. 1979). In the Smoky study, there were four leukemia cases among the $0.18 \times 0.76 = 15\%$ of cohort members who referred themselves and four among the $0.82 \times 0.76 = 62\%$ of cohort members traced by the investigators, for a total of eight cases among the 76% of the cohort with known outcomes. These data indicate that self-selection bias was a small but real problem in the Smoky study. If the 24% of the cohort with unknown outcomes had a leukemia incidence like that of the subjects traced by the investigators, we should expect that only $4(24/62) = 1.5$ or about 1 or 2 cases occurred among this 24%, for a total of only 9 or 10 cases in the entire cohort. If, however, we assumed that the 24% with unknown outcomes had a leukemia incidence like that of subjects with known outcomes, we would calculate that $8(24/76) = 2.5$ or about 2 or 3 cases occurred among this 24%, for a total of 10 or 11 cases in the entire cohort.

Self-selection can also occur before subjects are identified for study. For example, it is routine to find that the mortality of active workers is less than that of the population as a whole (McMichael 1976; Fox and Collier 1976). This “healthy-worker effect” presumably derives from a screening process, perhaps largely self-selection, that allows relatively healthy people to become or remain workers, whereas those who remain unemployed, retired, disabled, or otherwise out of the active worker population are as a group less healthy (Wang and Miettinen 1982).

3.6.2 Diagnostic Bias

Another type of selection bias occurring before subjects are identified for study is diagnostic bias (Sackett 1979). When the relation between oral contraceptives and venous thromboembolism was first investigated with case-control studies of hospitalized patients, there was a concern that some of the women had been hospitalized with a diagnosis of venous thromboembolism because their physicians suspected a relation between this disease and oral contraceptives and had known about oral contraceptive use in patients who presented with suggestive symptoms (Sartwell et al. 1969). A study of hospitalized patients with thromboembolism could lead to an exaggerated estimate of the effect of oral contraceptives on thromboembolism if the hospitalization and determination of the diagnosis were influenced by the history of oral contraceptive use.

3.7 Information Bias

Once the subjects to be compared have been identified, the information to be compared must be obtained. Bias in evaluating an effect can occur from errors in obtaining the needed information. Information bias can occur whenever there are errors in the measurement of subjects, but the consequences of the errors are different depending on whether the distribution of errors for one variable (e.g., exposure or disease) depends on the actual value of other variables.

For discrete variables (variables with only a countable number of possible values, such as indicators for sex), measurement error is usually called classification error or misclassification. Classification error that depends on the values of other variables is referred to as *differential misclassification*. Classification error that does not depend on the values of other variables is referred to as *non-differential misclassification*.

3.7.1 Differential Misclassification

Suppose a cohort study were undertaken to compare incidence rates of emphysema among smokers and non-smokers. Emphysema is a disease that may go undiagnosed without unusual medical attention. If smokers, because of concern about health-related effects of smoking or as a consequence of other health effects of smoking (such as bronchitis), seek medical attention to a greater degree than non-smokers, then emphysema might be diagnosed more frequently among smokers than among non-smokers simply as a consequence of the greater medical attention. Unless steps were taken to ensure comparable follow-up, an information bias would result: A spurious excess of emphysema incidence would be found among smokers compared with non-smokers that is unrelated to any biological effect of smoking. This is an example of differential misclassification, since the underdiagnosis of emphysema, a classification error, occurs more frequently for non-smokers than for smokers. Sackett (1979) has described it as a diagnostic bias, but unlike the diagnostic bias in the studies of oral contraceptives and thromboembolism described earlier, it is not a selection bias, since it occurs among subjects already included in the study. Nevertheless, the similarities between some selection biases and differential misclassification biases are worth noting.

In case-control studies of congenital malformations, the etiological information may be obtained at interview from mothers. The case mothers have recently given birth to a malformed baby, whereas the vast majority of control mothers have recently given birth to an apparently healthy baby. Another variety of differential misclassification, referred to as recall bias, can result if the mothers of malformed infants recall exposures more thoroughly than mothers of healthy infants. It is supposed that the birth of a malformed infant serves as a stimulus to a mother to recall all events that might have played some role in the unfortunate outcome. Presumably, such women will remember exposures such as infectious disease, trauma, and drugs more accurately than mothers of healthy infants, who have not had a comparable stimulus. Consequently, information on such exposures will be ascertained more frequently from mothers of malformed babies, and an apparent effect, unrelated to any biological effect, will result from this recall bias. Recall bias is a possibility in any case-control study that uses an anamnestic response, since the cases and controls by definition are people who differ with respect to their disease experience, and this difference may affect recall. Klemetti and Saxen (1967) found that the amount of time lapsed between the exposure and the recall was an important indicator of the accuracy of recall; studies in which the average time since exposure was different for interviewed cases and controls could thus suffer a differential misclassification.

The bias that is caused by differential misclassification can either exaggerate or underestimate an effect. In each of the examples above, the misclassification serves to exaggerate the effects under study, but examples to the contrary can also be found. Because of the relatively unpredictable effects of differential misclassification, some investigators go through elaborate procedures to ensure that the misclassification will be non-differential, such as blinding of exposure evaluations with respect to outcome status. Unfortunately, even in situations when blinding is accomplished or in cohort studies in which disease outcomes have not yet occurred, collapsing continuous or categorical exposure data into fewer categories can induce differential misclassification (Wacholder 1991; Flegal et al. 1991).

3.7.2 Non-Differential Misclassification

Non-differential exposure or disease misclassification occurs when the proportion of subjects misclassified on exposure does not depend on disease status or when the proportion of subjects misclassified on disease does not depend on exposure. Under certain conditions, any bias introduced by such non-differential misclassification of a binary exposure or disease is predictable in direction, namely, toward the null value (Newell 1962; Keys and Kihlberg 1963; Gullen et al. 1968; Copeland et al. 1977). Contrary to popular misconceptions, however, non-differential exposure or disease misclassification can sometimes produce bias away from the null (Walker and Blettner 1985; Dosemeci et al. 1990; Chavance et al. 1992; Kristensen 1992). For example, when both exposure and disease are non-differentially misclassified but the classification errors are dependent, it is possible to obtain bias away from the null (Chavance et al. 1992; Kristensen 1992), and the simple bias relations just given will no longer apply. Dependent errors can arise easily in many situations, such as in studies in which exposure and disease status are both determined from interviews.

Because the bias from independent non-differential misclassification of a dichotomous exposure is always in the direction of the null value, historically it has not been a great source of concern to epidemiologists, who have generally considered it more acceptable to underestimate effects than to overestimate effects. Nevertheless, such misclassification is a serious problem: The bias it introduces may account for certain discrepancies among epidemiological studies. Many studies ascertain information in a way that guarantees substantial misclassification, and many studies use classification schemes that can mask effects in a manner identical to non-differential misclassification.

Suppose aspirin transiently reduces risk of myocardial infarction. The word transiently implies a brief induction period. Any study that considered as exposure aspirin use outside of a narrow time interval before the occurrence of a myocardial infarction would be misclassifying aspirin use: There is relevant use of aspirin, and there is use of aspirin that is irrelevant because it does not allow the exposure to act causally under the causal hypothesis with its specified induction period. Many studies ask about "ever use" (use at any time during an individual's life) of drugs or other exposures. Such cumulative indices over an individual's lifetime inevitably augment possibly relevant exposure with irrelevant exposure and can thus introduce a bias toward the null value through non-differential misclassification.

Note that for a binary exposure variable with non-differential misclassification and independent classification errors, the direction of bias will be toward the null value, but if the exposure variable has more than two categories, one or more of those categories may manifest bias away from the null (Dosemeci et al. 1990). For example, suppose exposure had three levels, 0, 1, and 2, with risk increasing as level increased, and we wished to measure the effect of exposure at levels 1 and 2 relative to 0. If there was misclassification only between levels 1 and 2, the measures of effect for these two levels of exposure would be biased toward each other, and therefore the estimation for level 1 versus 0 would be biased away from the null (Walker and Blettner 1985).

In cohort studies in which there are disease categories with few subjects, investigators are occasionally tempted to combine outcome categories to increase the number of subjects in each analysis, thereby gaining precision. This collapsing of categories can obscure effects on more narrowly defined disease categories. For example, Smithells and Shepard (1978) investigated the teratogenicity of the drug Bendectin, a drug indicated for nausea of pregnancy. Because only 35 babies in their cohort study were born with a malformation, their analysis was focused on the single outcome, "malformation." But no teratogen causes all malformations; if such an analysis fails to find an effect, the failure may simply be the result of the grouping of many malformations not related to Bendectin with those that are. In fact, despite the authors' claim that "their study provides substantial evidence that Bendectin is not teratogenic in man," their data indicated a strong (though imprecise) relation between Bendectin and cardiac malformations. Unwarranted assurances of a lack of effect can easily emerge from studies in which a wide range of etiologically unrelated outcomes are grouped.

Non-differential exposure and disease misclassification is a greater concern in interpreting studies that seem to indicate the absence of an effect. Consequently, in studies that indicate little or no effect, it is crucial for the researchers to consider the problem of non-differential misclassification to determine to what extent a real effect might have been obscured. On the other hand, in studies that describe a strong non-zero effect, preoccupation with non-differential exposure and disease misclassification is rarely warranted, provided that the errors are independent.

Occasionally, critics of a study will argue that poor exposure data or a poor disease classification invalidates the results. This argument is incorrect, however, if the results indicate a non-zero effect and one can be sure that the classification errors produced bias toward the null, since the bias will be in the direction of underestimating the effect.

The importance of appreciating the likely direction of bias was illustrated by the interpretation of a study on spermicides and birth defects (Jick et al. 1981a, b). This study reported an increased prevalence of several types of congenital disorder among women who were identified as having filled a prescription for spermicides during a specified interval before the birth. The exposure information was only a rough correlate of the actual use of spermicides during a theoretically relevant time period, but the misclassification that resulted was in all probability non-differential and independent of errors in outcome ascertainment, because prescription information was recorded on a computer log before the outcome was known. One of the criticisms raised about the study was that inaccuracies in the exposure information cast doubt on the validity of the findings (Felarca et al. 1981; Oakley 1982). Whatever bias was present on this account, however, would not likely have led to an underestimation of any real effect, so this criticism is inappropriate (Jick et al. 1981b).

Generally speaking, it is incorrect to dismiss a study reporting an effect simply because there is substantial non-differential misclassification of exposure, since an estimate of effect without the misclassification could be even greater, provided that the misclassification probabilities apply uniformly to all subjects. Thus, the implications of non-differential misclassification depend heavily on whether the study is perceived as “positive” or “negative.” Emphasis on measurement instead of on a qualitative description of study results lessens the likelihood for misinterpretation, but even so, it is important to bear in mind the direction and likely magnitude of a bias.

3.7.3 Misclassification of Confounders

If a confounding variable is misclassified, the ability to control confounding in the analysis is hampered (Greenland 1980; Savitz and Baron 1989; Brenner 1993; Marshall and Hastrup 1996; Marshall et al. 1999). While independent non-differential misclassification of exposure or disease usually biases study results in the direction of the null hypothesis, independent non-differential misclassification of a confounding variable will usually reduce the degree to which confounding can be controlled and thus can cause a bias in either direction, depending on the direction of the confounding. For this reason, misclassification of confounding factors can be a serious problem.

If the confounding is strong and the exposure-disease relation is weak or zero, misclassification of the confounding factor can lead to extremely misleading results. For example, a strong causal relation between smoking and bladder cancer, coupled with a strong association between smoking and coffee drinking, makes smoking a strong confounder of any possible relation between coffee drinking and bladder cancer. Since the control of confounding by smoking depends on accurate smoking information, and since some misclassification of the relevant smoking information is inevitable no matter how smoking is measured, some residual confounding is inevitable (Morrison et al. 1982). The problem of residual confounding would be even worse if the only available information on smoking were a simple dichotomy such as "ever smoked" versus "never smoked," since the lack of detailed specification of smoking prohibits adequate control of confounding. The resulting confounding is especially troublesome because to many investigators and readers, it may appear that confounding by smoking has been controlled.

3.8 Conclusion

Epidemiology is concerned with making inferences about the distribution and causes of disease and health in human populations. One should bear in mind that these inferences, like any scientific inference, can never be drawn with complete certainty and will often be highly tentative in light of unresolved validity issues, such as uncontrolled confounding. The uncertainties stemming from validity issues cannot always be addressed by statistical methods; hence, the process of epidemiological inference is a more complicated process than statistical inference. Epidemiological inference is further complicated by subtleties that arise when quantifying and measuring population effects, such as the distinction between number of individuals harmed by an exposure and the excess caseload produced by an exposure. These subtleties also cannot be addressed using ordinary statistical theory, and yet they can be of crucial importance in attempts to employ epidemiological results in decision-making contexts. The proper conduct and interpretation of epidemiological research and its application in public health requires mastery of epidemiological concepts and methods that are outlined in this chapter and elucidated further in the subsequent chapters of this handbook.

References

Alho JM (1992) On prevalence, incidence and duration in stable populations. Biometrics 48: 578–592

Bayes T (1763) Essay towards solving a problem in the doctrine of chances. Philos Trans R Soc 53:370–418

Brenner H (1993) Bias due to non-differential misclassification of polytomous confounders. J Clin Epidemiol 46:57–63

Caldwell GG, Kelley DB, Heath CW Jr (1980) Leukemia among participants in military maneuvers at a nuclear bomb test: a preliminary report. JAMA 244:1575–1578

Chavance M, Dellatolas G, Lellouch J (1992) Correlated nondifferential misclassifications of disease and exposure. Int J Epidemiol 21:537–546

Cole P, MacMahon B (1971) Attributable risk percent in case-control studies. Br J Prev Soc Med 25:242–244

Copeland KT, Checkoway H, Holbrook RH, McMichael AJ (1977) Bias due to misclassification in the estimate of relative risk. Am J Epidemiol 105:488–495

Cornfield J (1976) Recent methodological contributions to clinical trials. Am J Epidemiol 104:408–424

Criqui MH, Austin M, Barrett-Connor E (1979) The effect of non-response on risk ratios in a cardiovascular disease study. J Chronic Dis 32:633–638

Curd M, Cover JA (eds) (1998) Philosophy of science, section 3: the duhem-quine thesis and underdetermination. W.W. Norton & Company, New York

DeFinetti B (1937) Foresight: its logical laws, its subjective sources. Reprinted in: Kyburg HE, Smokler HE (eds) (1964) Studies in subjective probability. Wiley, New York

Deubner DC, Wilkinson WE, Helms MJ, Tyroler HA, Hanes CG (1980) Logistic model estimation of death attributable to risk factors for cardiovascular disease in Evans County, Georgia. Am J Epidemiol 112:135–143

Doll R, Peto R (1981) The causes of cancer. Oxford University Press, New York

Dosemeci M, Wacholder S, Lubin J (1990) Does nondifferential misclassification of exposure always bias a true effect toward the null value? Am J Epidemiol 132:746–749

Duhem P (1906) La théorie physique son objet et sa structure (The aim and structure of physical theory) (Trans: from the French by Wiener PP, 1954), Princeton University Press, Princeton

Felarca LC, Wardell DM, Rowles B (1981) Vaginal spermicides and congenital disorders. JAMA 246:2677

Flegal KM, Keyl PM, Nieto FJ (1991) Differential misclassification arising from nondifferential errors in exposure measurement. Am J Epidemiol 134:1233–1244

Fox AJ, Collier PF (1976) Low mortality rates in industrial cohort studies due to selection for work and survival in the industry. Br J Prev Soc Med 30:225–230

Freeman J, Hutichison GB (1980) Prevalence, incidence and duration. Am J Epidemiol 112:707–723

Gilovich T, Griffin D, Kahneman D (2002) Heuristics and biases: the psychology of intuitive judgment. Cambridge University Press, NewYork

Greenland S (1980) The effect of misclassification in the presence of covariates. Am J Epidemiol 112:564–569

Greenland S (1987) Interpretation and choice of effect measures in epidemiologic analysis. Am J Epidemiol 125:761–768

Greenland S (1999) The relation of the probability of causation to the relative risk and the doubling dose: a methodologic error that has become a social problem. Am J Public Health 89: 1166–1169

Greenland S (2002) Causality theory for policy uses of epidemiologic measures. Ch. 6.2 In: Murray CJL, Salomon JA, Mathers CD, Lopez AD (eds) Summary measures of population health. Harvard University Press/WHO, Cambridge, pp 291–302

Greenland S (2005) Epidemiologic measures and policy formulation: lessons from potential outcomes (with discussion). Emerg Themes Epidemiol 2:1–4

Greenland S, Brumback BA (2002) An overview of relations among causal modelling methods. Int J Epidemiol 31:1030–1037

Greenland S, Robins JM (1986) Identifiability, exchangeability and epidemiological confounding. Int J Epidemiol 15:413–419

Greenland S, Robins JM (1988) Conceptual problems in the definition and interpretation of attributable fractions. Am J Epidemiol 128:1185–1197

Greenland S, Robins JM (2000) Epidemiology, justice, and the probability of causation. Jurimetrics 40:321–340

Greenland S, Robins JM, Pearl J (1999) Confounding and collapsibility in causal inference. Stat Sci 14:19–46

Gullen WH, Berman JE, Johnson EA (1968) Effects of misclassification in epidemiologic studies. Public Health Rep 53:1956–1965

Haack S (2003) Defending science – within reason. Between scientism and cynicism. Prometheus Books, Amherst

Hernán MA (2005) Hypothetical interventions to define causal effects – afterthought or prerequisite? Am J Epidemiol 162:618–620

Hill AB (1965) The environment and disease: association or causation? Proc R Soc Med 58: 295–300

Horwitz RI, Feinstein AR (1978) Alternative analytic methods for case-control studies of estrogens and endometrial cancer. N Engl J Med 299:1089–1094

Howson C, Urbach P (1993) Scientific reasoning: the bayesian approach, 2nd edn. Open Court, LaSalle

Jick H, Walker AM, Rothman KJ, Hunter JR, Holmes LB, Watkins RN, D'Ewart DC, Danford A, Madsen S (1981a) Vaginal spermicides and congenital disorders. JAMA 245:1329–1332

Jick H, Walker AM, Rothman KJ, Holmes LB (1981b) Vaginal spermicides and congenital disorders - reply. JAMA 246:2677–2678

Kahnemann D, Slovic P, Tversky A (1982) Judgment under uncertainty: heuristics and biases. Cambridge University Press, New York

Keiding N (1991) Age-specific incidence and prevalence: a statistical perspective. J R Stat Soc A 154:371–412

Keys A, Kihlberg JK (1963) The effect of misclassification on the estimated relative prevalence of a characteristic. Am J Public Health 53:1656–1665

Klemetti A, Saxen L (1967) Prospective versus retrospective approach in the search for environmental causes of malformations. Am J Public Health 57:2071–2075

Koepsell TD, Weiss NS (2003) Epidemiologic methods: studying the occurrence of illness. Oxford University Press, New York

Kristensen P (1992) Bias from nondifferential but dependent misclassification of exposure and outcome. Epidemiology 3:210–215

Lanes SF, Poole C (1984) "Truth in packaging?" the unwrapping of epidemiologic research. J Occup Med 26:571–574

Levin ML (1953) The occurrence of lung cancer in man. Acta Unio Int Contra Cancrum 9:531–541

Lewis D (1973) Causation. J Philos 70:556–567. Reprinted with postscript in: Lewis D (1986) Philosophical papers. Oxford, New York

Mackie J (1965) Causes and conditions. Am Philos Q 2:245–255. Reprinted in: Sosa E, Tooley M (eds) (1993) Causation. Oxford, New York, pp 33–55

Maclure M (1985) Popperian refutation in epidemiology. Am J Epidemiol 121:343–350

MacMahon B, Pugh TF (1967) Causes and entities of disease. In: Clark DW, MacMahon B (eds) Preventive medicine. Little, Brown and Co., Boston

MacMahon B, Pugh TF (1970) Epidemiology: principles and methods. Little, Brown and Co., Boston, pp 137–198, 175–184

Magee B (1985) Philosophy and the real world. An introduction to karl popper. Open Court, La Salle

Marshall JR, Hastrup JL (1996) Mismeasurement and the resonance of strong confounders: uncorrelated errors. Am J Epidemiol 143:1069–1078

Marshall JR, Hastrup JL, Ross JS (1999) Mismeasurement and the resonance of strong confounders: correlated errors. Am J Epidemiol 150:88–96

McMichael AJ (1976) Standardized mortality ratios and the "healthy worker effect": scratching beneath the surface. J Occup Med 18:165–168

Medawar PB (1979) Advice to a young scientist. Basic Books, New York

Miettinen OS (1976) Estimability and estimation in case-referent studies. Am J Epidemiol 103:226–235

Miettinen OS, Cook EF (1981) Confounding: essence and detection. Am J Epidemiol 114:593–603

Mill JS (1843) A system of logic, ratiocinative and inductive, 5th edn. Parker, Son and Bowin, London

Morgenstern H, Bursic ES (1982) A method for using epidemiologic data to estimate the potential impact of an intervention on the health status of a target population. J Community Health 7:292–309

Morrison AS (1979) Sequential pathogenic components of rates. Am J Epidemiol 109:709–718

Morrison AS, Buring JE, Verhoek WG, Aoki K, Leck I, Ohno Y, Obata K (1982) Coffee drinking and cancer of the lower urinary tract. J Natl Cancer Inst 68:91–94

Newell DJ (1962) Errors in interpretation of errors in epidemiology. Am J Public Health 52: 1925–1928

Oakley G Jr (1982) Spermicides and birth defects. JAMA 247:2405

Ouellet BL, Ræmeder J-M, Lance J-M (1979) Premature mortality attributable to smoking and hazardous drinking in Canada. Am J Epidemiol 109:451–463

Peacock PB (1971) The non-comparability of relative risks from different studies. Biometrics 27:903–907

Pearce NE (1990) White swans, black ravens, and lame ducks: necessary and sufficient causes in epidemiology. Epidemiology 1:47–50

Pearl J (2009) Causality, 2nd ed. Cambridge University Press, Cambridge

Phillips CV, Goodman KJ (2004) The missed lessons of Sir Austin Bradford Hill. Epidemiol Perspect Innov 1:3. doi:10.1186/1742-5573-1-3

Platt JR (1964) Strong inference. Science 146:347–353

Popper KR (1934) Logik der Forschung. Julius Springer, Vienna; English translation (1959) The Logic of Scientific Discovery, Basic Books, New York

Popper KR (1959) The propensity interpretation of probability. Brit J Phil Sci, 10:25–42

Preston SH (1987) Relations among standard epidemiologic measures in a population. Am J Epidemiol 126:336–345

Quine WVO (1951) Two dogmas of empiricism. Philos Rev 60:20–43. Reprinted with edits in: Quine WVO (1953) From a logical point of view. Harvard University Press (2nd revised ed 1961)

Ramsey FP (1931) Truth and probability. Reprinted in: Kyburg HE, Smokler HE (eds) Studies in subjective probability. Wiley, New York, 1964

Rothman KJ (1976) Causes. Am J Epidemiol 104:587–592

Rothman KJ (1981) Induction and latent periods. Am J Epidemiol 114:253–259

Rothman KJ (ed) (1988) Causal inference. Epidemiology Resources, Inc., Boston

Rothman KJ, Greenland S, Lash TL (eds) (2008) Modern epidemiology, 3rd ed. Lippincott-Williams-Wilkins, Philadelphia

Rubin DB (1990) Comment: Neyman (1923) and causal inference in experiments and observational studies. Stat Sci 5:472–480

Sackett DL (1979) Bias in analytic research. J Chron Dis 32:51–63

Sartwell PE, Masi AT, Arthes FG, Greene GR, Smith HE (1969) Thromboembolism and oral contraceptives: an epidemiologic case-control study. Am J Epidemiol 90:365–380

Savitz DA, Baron AE (1989) Estimating and correcting for confounder misclassification. Am J Epidemiol 129:1062–1071

Smithells RW, Shepard S (1978) Teratogenicity testing in humans: a method demonstrating the safety of Bendectin. Teratology 17:31–36

Susser M (1973) Causal thinking in the health sciences. Oxford, New York

Susser M (1991) What is a cause and how do we know one? A grammar for pragmatic epidemiology. Am J Epidemiol 133:635–648

Taubes G (1993) Bad science. The short life and weird times of cold fusion. Random House, New York

U.S. Department of Health, Education and Welfare. Smoking and Health (1964) Report of the advisory committee to the surgeon general of the public health service. Public Health Service Publication No. 1103. Government Printing Office, Washington, DC

Wacholder S (1991) Practical considerations in choosing between the case-cohort and nested case-control design. Epidemiology 2:155–158

Walker AM, Blettner M (1985) Comparing imperfect measures of exposure. Am J Epidemiol 121:783–790

Wang J, Miettinen OS (1982) Occupational mortality studies: principles of validity. Scand J Work Environ Health 8:153–158

Weed D (1986) On the logic of causal inference. Am J Epidemiol 123:965–979

Weiss NS (2002) Can the specificity of an association be rehabilitated as a basis for supporting a causal hypothesis? Epidemiology 13:6–8

Rates, Risks, Measures of Association and Impact

4

Jacques Benichou and Mari Palta

Contents

J. Benichou (✉)
Department of Biostatistics, University of Rouen Medical School and Rouen University Hospital,
Rouen Cedex, France

M. Palta
Department of Population Health Sciences and Department of Biostatistics and Medical
Informatics, University of Wisconsin School of Medicine and Public Health, Wisconsin,
WI, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_3,

4.1 Introduction

A major aim of epidemiological research is to measure disease occurrence in relation to various characteristics such as exposure to environmental, occupational, or lifestyle risk factors, genetic traits, or other features. In this chapter, various measures will be considered that quantify disease occurrence, associations between disease occurrence and these characteristics, as well as their consequences in terms of both disease risk and impact at the population level. As a common practice, the generic term exposure will be used throughout the chapter to denote such characteristics. Emphasis will be placed on measures based on occurrence of new disease cases, referred to as disease incidence. Measures based on disease prevalence, i.e., considering newly occurring and previously existing disease cases as a whole, will be considered more briefly.

We will first define the basic measure of disease incidence, namely, the incidence rate, from which other measures considered in this chapter can be derived. These other measures, namely, measures of disease risk, measures of association between exposure and disease risk (e.g., relative risk), and measures of impact of exposure-disease associations (e.g., attributable risk), will be considered successively. Additional points will be made regarding standardized incidence rates and measures based on prevalence.

4.2 Incidence and Hazard Rates

4.2.1 Definition

The incidence rate of a given disease is the number of persons observed to develop the disease (number of incident cases) among subjects at risk of developing the disease in the source population over a chosen time period or age interval. Incidence rates are not interpretable as probabilities. While they have a lower bound of zero, they have no upper bound. Units of incidence rates are reciprocals of person-time, such as reciprocals of person-years or multiples of person-years (e.g., 100,000 person-years). For instance, if 10 cases develop from the follow-up of 20 subjects and for a total follow-up time of 5 years, the incidence rate is $10/100 = 0.1$ cases per person-year (assuming an instantaneous event with immediate recovery and all 20 subjects being at risk until the end of the observation period).

Usually, incidence rates are assessed over relatively short time periods compared with the time scale for disease development, e.g., intervals of 5-years for chronic diseases with an extended period of susceptibility such as many cancers.

Synonyms for incidence rate are average incidence rate, force of morbidity, person-time rate, or incidence density (Miettinen 1976) with the last term reflecting the interpretation of an incidence rate as the density of incident case occurrences in an accumulated amount of person-time (Morgenstern et al. 1980). Mortality rates (overall or cause-specific) can be regarded as special cases of incidence rates, the outcome considered being death rather than disease occurrence.

Incidence rates can be regarded as estimates of a limiting theoretical quantity, namely, the hazard rate, $h(t)$, also called the incidence intensity or force of morbidity. The hazard rate at time t, $h(t)$, is the instantaneous rate of developing the disease of interest in an arbitrarily short interval Δ around time t, provided the subject is still at risk at time t (i.e., has not fallen ill before time t). Technically, it has the following mathematical definition:

$$h(t) = \text{limit}_{\Delta \downarrow 0} \Delta^{-1} \Pr(t \le T < t + \Delta | t \le T), \tag{4.1}$$

where T is the time period for the development of the disease considered and Pr denotes probability. Indeed, for time intervals in which the hazard rate can be assumed constant, the incidence rate as defined above represents a valid estimate of the hazard rate. Thus, this result applies when piecewise constant hazard rates are assumed, which can be regarded as realistic in many applications, especially when reasonably short time intervals are used, and leads to convenient estimating procedures, e.g., based on the Poisson model.

Strictly speaking, incidence and hazard rates do not coincide. Hazard rates are formally defined as theoretical functions of time whereas incidence rates are defined directly as estimates and constitute valid estimates of hazard rates under certain assumptions (see above). For the sake of simplicity, however, we will use the terms incidence rates and hazard rates as synonyms in the remainder of this chapter unless a clear distinction is needed.

4.2.2 Estimability and Basic Principles of Estimation

From the definitions above, it ensues that individual follow-up data are needed to obtain incidence rates or estimate hazard rates. Alternatively, in the absence of individual follow-up data, person-time at risk can be estimated as the time period width times the population size at midpoint. Such estimation makes the assumption that individuals who disappear from being at risk, either because they succumb or because they move in or out, do so evenly across the time interval. Thus, population data such as registry data can be used to estimate incidence rates as long as an exhaustive census of incident cases can be obtained.

Among the main designs considered in Part I (Concepts and Designs in Epidemiology) of this handbook, the cohort design (cf. chapter ▸Cohort Studies of this handbook) is the ideal design to obtain incidence or hazard rates for various levels or profiles of exposure, i.e., exposure-specific incidence or hazard rates. This is because follow-up is available on subjects with various profiles of exposure. In many applications, obtaining exposure-specific incidence rates is not trivial, however. Indeed, multiple exposures are often considered, some with several exposed levels and some continuous. Moreover, it may be necessary to account for confounders or effect modifiers. Hence, estimation often requires modeling. Methods of inference based on regression models are considered in detail in Part III (Statistical Methods in Epidemiology) of this handbook, particularly chapter ▶Regression Methods for Epidemiological Analysis.

Case-control data (cf. chapter ▶Case-Control Studies of this handbook) pose a more difficult problem than cohort data because case-control data alone are not sufficient to yield incidence or hazard rates. Indeed, they provide data on the distributions of exposure respectively in diseased subjects (cases) and non-diseased subjects (controls) for the disease under study, which can be used to estimate odds ratios (see Sect. 4.4) but are not sufficient to estimate exposure-specific incidence rates. However, it is possible to arrive at exposure-specific incidence rates from case-control data if case-control data are complemented by either follow-up or population data, which happens for nested or population-based case-control studies. In a nested case-control study, the cases and controls are selected from a follow-up study. In a population-based case-control study, they are selected from a specified population in which an effort is made to identify all incident cases diagnosed during a fixed time interval, usually in a grouped form (e.g., number of cases and number of subjects by age group). In both situations, full information on exposure is obtained only for cases and controls. Additionally, complementary information on composite incidence (i.e., counts of events and person-time) can be sought from the follow-up or population data. By combining this information with odds ratio estimates, exposure-specific incidence rates can be obtained. This has long been recognized (Cornfield 1951, 1956; MacMahon 1962; Miettinen 1974, 1976; Neutra and Drolette 1978) and is a consequence of the relation (Miettinen 1974; Gail et al. 1989):

$$h_0 = h^*(1 - AR), \tag{4.2}$$

where AR is the attributable risk in the population for all exposures considered, a quantity estimable from case-control data (see Sect. 4.5); h_0 is the baseline incidence rate, i.e., the incidence rate for subjects at the reference (unexposed) level of all exposures considered; and h^* is the composite or average incidence rate in the population that includes unexposed subjects and subjects at various levels of all exposures (i.e., with various profiles of exposure). The composite incidence rate h^* can be estimated from the complementary follow-up or population data. Equation 4.2 simply states that the incidence rate for unexposed subjects is equal to the proportion of the average incidence rate in the population that is not associated with any of the exposures considered. Equation 4.2 can be specialized to various subgroups or strata defined by categories of age, sex, or geographical location such as region or center, on which incidence rates are assumed constant. From the baseline rate h_0, incidence rates for all levels or profiles of exposure can be derived using odds ratio estimates, provided odds ratio estimates are reasonable estimates of incidence rate ratios as in the case of a rare disease (see Sect. 4.4). Consequently, exposure-specific incidence rates can be obtained from case-control data as long as they are complemented by follow-up or population data that can be used to estimate average incidence rates.

Example 4.1.
Exposure-specific incidence rates of breast cancer were obtained based on age as well as family history in first-degree relatives, reproductive history (i.e., age at menarche and age at first live birth), and history of benign disease from the Breast Cancer Detection and Demonstration Project (BCDDP). The BCDDP combined the prospective follow-up of 284,780 women over 5 years and a nested case-control study (Gail et al. 1989) with about 3,000 cases and 3,000 controls. For each 5-year age group from ages 35 to 79 years, composite incidence rates were obtained from the follow-up data. In age groups 40–44 and 45–49 years, 162 and 249 new cases of breast cancer developed from the follow-up of 79,526.4 and 88,660.7 person-years, yielding composite incidence rates of 203.7 and 280.8 per 10^5 person-years, respectively. For all women less than 50 years of age, the attributable risk for family history, reproductive history, and history of benign breast disease was estimated at 0.4771 from the nested case-control data (see Sect. 4.5). By applying Eq. 4.2, baseline incidence rates for women at the reference level of all these factors were $203.7 \times (1 - 0.4771) = 106.5$ and $280.8 \times (1 - 0.4771) = 146.8$ per 10^5 person-years, respectively. For a nulliparous woman of age 40, with menarche at age 12, one previous biopsy for benign breast disease, and no history of breast-cancer in her first-degree relatives, the corresponding odds ratio was estimated at 2.89 from logistic regression analysis of the nested case-control data (see Sect. 4.4.6.3), yielding an exposure-specific incidence rate of $106.5 \times 2.89 = 307.8$ per 10^5 person-years. For a 45-year-old woman with the same exposure profile, the corresponding exposure-specific incidence rate was $146.8 \times 2.89 = 424.3$ per 10^5 person-years.

Finally, cross-sectional data cannot provide any assessment of incidence rates but instead will yield estimates of disease prevalence proportions as discussed in Sect. 4.6 of this chapter.

4.2.3 Relation with Other Measures

The reason why exposure-specific incidence or hazard rates are central quantities is that, once they are available, most other quantities described in this chapter can be obtained from them, namely, measures of disease risk, measures of association between exposure and disease risk, and measures of exposure impact in terms of new disease burden at the population level. However, it should be noted that measures of impact as well as some measures of association (i.e., odds ratios) can be estimated from case-control data alone without relying on exposure-specific incidence rates (see Sects. 4.3 and 4.4). Moreover, cross-sectional data can yield estimates of measures of association and impact with respect to disease prevalence (see Sect. 4.6).

4.3 Measures of Disease Risk

4.3.1 Definition

Disease risk is defined as the probability that an individual who is initially disease-free will develop a given disease over a specified time or age interval (e.g., 1 year or lifetime). Of all incidence and risk measures, this measure is probably the one most familiar and interpretable to consumers of health data.

If the interval starting at time a_1 and ending just before time a_2, i.e., $[a_1, a_2)$, is considered, disease risk can be written formally as

$$\pi(a_1, a_2) = \int_{a_1}^{a_2} h(a)\{S(a)/S(a_1)\}da. \tag{4.3}$$

In Eq. 4.3, $h(a)$ denotes the disease hazard rate at time or age a (see Sect. 4.2). The function $S(\cdot)$, with $(\cdot)$ an arbitrary argument, is the survival function, so that $S(a)$ denotes the probability of still being disease-free at time at age a and $S(a)/S(a_1)$ denotes the conditional probability of staying disease-free up to time or age a for an individual who is free of disease at the beginning of the interval $[a_1, a_2)$. Equation 4.3 integrates over the interval $[a_1, a_2)$ the instantaneous incidence rate of developing disease at time or age a for subjects still at risk of developing the disease (i.e., subjects still disease-free). Because the survival function $S(\cdot)$ can be written as a function of the disease hazard rate through

$$S(a_2)/S(a_1) = \exp\left\{-\int_{a_1}^{a_2} h(a)da\right\}, \tag{4.4}$$

disease risk is also a function of the disease hazard rate.

By specifying the functions $h(\cdot)$ and $S(\cdot)$, various quantities can be obtained that measure disease risk in different contexts. First, the time scale on which these functions as well as disease risk are defined corresponds to two specific uses of risk. In most applications, the relevant time scale is age, since disease incidence is influenced by age in most applications. Note that by considering the age interval $[0, a_2)$, one obtains lifetime disease risk up to age a_2. However, in clinical epidemiology settings, risk refers to the occurrence of an event, such as relapse or death in subjects already presenting with the disease of interest. In this context, the relevant time scale becomes time from disease diagnosis or, possibly, time from some other disease-related event, such as a surgical resection of a tumor or occurrence of a first myocardial infarction.

Second, risk definition may or may not account for individual exposure profiles. If no risk factors are considered in the estimation of the disease hazard rate, the corresponding measure of disease risk defines the average or composite risk over the entire population that includes subjects with various exposure profiles. This measure, also called cumulative incidence (Miettinen 1976), may be of value at the population level. However, the main usefulness of risk is in quantifying an individual's predicted probability of developing disease depending on the individual's exposure profile. Thus, estimates of exposure-specific disease hazard rate have to be available for such exposure-specific risk (also called individualized or absolute risk) to be estimated.

Third, the consideration of competing risks and the corresponding definition of the survival function $S(\cdot)$ yields two separate definitions of risk. Indeed, although risk is defined with respect to the occurrence of a given disease, subjects can die from other causes (i.e., competing risks), which obviously precludes disease occurrence. The first option is to define $S(a)$ as the theoretical probability of being disease-free at time or age a if other causes of death (competing risks) were eliminated yielding a measure of disease risk in a setting with no competing risks. This measure may not be of much practical value. Moreover, unless unverifiable assumptions regarding incidence of the disease of interest and deaths from other causes can be made, for instance, assuming that they occur independently, the function $S(\cdot)$ will not be estimable. For these reasons, it is more feasible to define $S(a)$ as the probability that an individual will be alive and disease-free at age a as the second option, yielding a more practical definition of disease risk as the probability of developing disease in the presence of competing causes of death (see Sect. 4.3.5).

From the definition of disease risk above, it appears that disease risk depends on the incidence rate of disease in the population considered and can also be influenced by the strength of the relationship between exposures and disease if individual risk is considered. One consequence is that risk estimates may not be portable from one population to another, as incidence rates may vary widely among populations that are separated in time and location or even among subgroups of populations, possibly because of differing genetic patterns or differing exposure to unknown risk factors. Additionally, competing causes of death (competing risks) may have different patterns among different populations, which might also influence values of disease risk.

4.3.2 Range

Disease risk is a probability and therefore lies between 0 and 1 and is dimensionless. A value of 0 while theoretically possible would correspond to very special cases such as a purely genetic disease for an individual not carrying the disease gene. A value of 1 would be even more unusual and might again correspond to a genetic disease with a penetrance of 1 for a gene carrier, but, even in this case, the value should be less than 1 if competing risks are accounted for.

4.3.3 Synonyms

Beside the term "disease risk," "absolute risk" or "absolute cause-specific risk" have been used by several authors (Dupont 1989; Benichou and Gail 1990a, 1995; Benichou 2000a; Langholz and Borgan 1997). Alternative terms include "individualized risk" (Gail et al. 1989), "individual risk" (Spiegelman et al. 1994), "crude probability" (Chiang 1968), "crude incidence" (Korn and Dorey 1992), "cumulative incidence" (Gray 1988; Miettinen 1976), "cumulative incidence risk" (Miettinen 1974), and "absolute incidence risk" (Miettinen 1976).

The term "cumulative risk" refers to the quantity $\int_{a_1}^{a_2} h(a)da$ and approximates disease risk closely in the case where disease is rare.

The term "attack rate" defines the risk of developing a communicable disease during a local outbreak and for the duration of the epidemic or the time during which primary cases occur (MacMahon and Pugh 1970, Chap. 5; Rothman and Greenland 1998, Chap. 27).

The term “floating absolute risk,” introduced by Easton et al. (1991), refers to a concept different from disease risk. It was derived to remedy the standard problem that measures of association such as ratios of rates, risks, or odds are estimated in reference to a baseline group, which causes their estimates for different levels of exposure to be correlated and may lead to lack of precision if the baseline group is small. The authors proposed a procedure to obtain estimates unaffected by these problems and used the term “floating absolute risk” to indicate that standard errors were not estimated in reference to an arbitrary baseline group.

4.3.4 Interpretation and Usefulness

Overall or average risks provide results such as “one in nine women will develop breast cancer at some time during her life” (American Cancer Society 1992), which are of no direct use in quantifying the risk of women with given exposure profiles and no direct help in deciding on preventive treatment or surveillance measures.

Upon taking individual exposure profiles into account, resulting individual disease risk estimates become useful in providing an individual measure of the probability of disease occurrence and can therefore be useful in counseling. They are well suited to predicting risk for an individual, unlike measures of association that quantify the increase in the probability of disease occurrence relative to subjects at the baseline level of exposure, but do not quantify that probability itself.

Individual risk has been used as a tool for individual counseling in breast cancer (Benichou et al. 1996; Gail and Benichou 1994; Hoskins et al. 1995). Indeed, a woman’s decision to take a preventive treatment such as tamoxifen (Fisher et al. 1998; Wu and Brown 2003) or even undergo prophylactic mastectomy (Hartman et al. 2001; Lynch et al. 2001) depends on her awareness of the medical options, on personal preferences, and on individual risk. A woman may have several risk factors, but if her individual risk of developing breast cancer over the next 10 years is small, she may be reassured and she may be well advised simply to embark on a program of surveillance. Conversely, she may be very concerned about her absolute risk over a longer time period, such as 30 years, and she may decide to use prophylactic medical treatment or even undergo prophylactic mastectomy if her absolute risk is very high.

Estimates of individual risk of breast cancer are available based on age, family history, reproductive history and history of benign disease (Gail et al. 1989; Costantino et al. 1999) and were originally derived from the BCDDP that combined a follow-up study and a nested case-control study (Gail et al. 1989). This example illustrates that not only exposures or risk factors per se (such as family history) may be used to obtain individual risk estimates but also markers of risk such as benign breast disease which are known to be associated with an increase in disease risk and may reflect some premalignant stage. In the same fashion, it has been suggested to improve existing individual risk estimates of breast cancer by incorporating mammographic density, a risk marker known to be associated with increased breast cancer risk (Benichou et al. 1997). In the cardiovascular field, individual risk estimates of developing myocardial infarction, developing coronary heart disease, dying from coronary heart disease, developing stroke, developing cardiovascular disease, and dying from cardiovascular disease were derived from the Framingham heart and Framingham offspring cohort studies. These estimates are based on age, sex, HDL, LDL, and total cholesterol levels, smoking status, blood pressure, and diabetes history (Anderson et al. 1991).

Individual risk is also useful in designing and interpreting trials of interventions to prevent the occurrence of a disease. At the design stage, disease risk may be used for sample size calculations because the sample sizes required for these studies depend importantly on the risk of developing the disease during the period of study and the expected distribution of exposure profiles in the study sample (Anderson et al. 1992). Disease risk has also been used to define eligibility criteria in such studies. For example, women were enrolled in a preventive trial to decide whether the drug tamoxifen can reduce the risk of developing breast cancer (Fisher et al. 1998). Because tamoxifen is a potentially toxic drug and because it was to be administered to a healthy population, it was decided to restrict eligibility to women with somewhat elevated absolute risks of breast cancer. All women over age 59 as well as younger women whose absolute risks were estimated to equal or exceed that of a typical 60-year-old woman were eligible to participate (Fisher et al. 1998). Individual risk has been used to interpret results of this trial through a risk-benefit analysis in order to help define which women are more likely to benefit from using tamoxifen. Women were identified who had a decrease in breast cancer risk and other events such as hip fracture from using tamoxifen surpassing the tamoxifen-induced increase in other events such as endometrial cancer, pulmonary embolism, or deep vein thrombosis (Gail et al. 1999).

Disease risk can also be important in decisions affecting public health. For example, in order to estimate the absolute reduction in lung cancer incidence that might result from measures to reduce exposure to radon, one could categorize a general population into subgroups based on age, sex, smoking, status and current radon exposure levels and then estimate the absolute reduction in lung cancer incidence that would result from lowering radon levels in each subgroup (Benichou and Gail 1990a; Gail 1975). Such an analysis would complement estimation of population attributable risk or generalized impact fractions (see Sect. 4.5).

The concept of risk is also useful in clinical epidemiology as a measure of the individualized probability of an adverse event, such as a recurrence or death in diseased subjects. In that context, risk depends on factors that are predictive of recurrence or death, rather than on factors influencing the risk of incident disease, and the time scale of interest is usually time from diagnosis or from surgery rather than age. It can serve as a useful tool to help define individual patient management, and, for instance, the absolute risk of recurrence in the next 3 years might be an important element in deciding whether to prescribe an aggressive and potentially toxic treatment regimen (Benichou and Gail 1990a; Korn and Dorey 1992).

4.3.5 Properties

Two main points need to be emphasized. First, as is evident from its definition, disease risk can only be estimated and interpreted in reference to a specified age or time interval. One might be interested in short time spans (e.g., 5 years) or long time spans (e.g., 30 years). Of course, disease risk increases as the time span increases. Sometimes, the time span is variable such as in lifetime risk.

Disease risk can be influenced strongly by the intensity of competing risks (typically competing causes of death, see above). Disease risk varies inversely as a function of death rates from other causes.

4.3.6 Estimability

It follows from its definition that disease risk is estimable as long as hazard rates for the disease (or event) of interest are estimable. Therefore, disease risk is directly estimable from cohort data, but case-control data have to be complemented with follow-up or population data in order to obtain the necessary complementary information on incidence rates (see Sect. 4.2).

It has been argued above (see Sect. 4.3.1) that disease risk is a more useful measure when it takes into account competing risks, i.e., the possibility for an individual to die of an unrelated disease before developing the disease (or disease-related event) of interest. In this setting, disease risk is defined as the probability of disease occurrence *in the presence* of competing risks, which is more relevant for individual predictions and other applications discussed above than the underlying (or "net" or "latent") probability of disease occurrence *in the absence* of competing risks. Moreover, disease risk is identifiable without any unverifiable competing risk assumptions in this setting, such as the assumption that competing risks act independently of the cause of interest because, as Prentice et al. (1978) emphasize, all functions of the disease hazard rates are estimable. Death rates from other causes can be estimated either internally from the study data or from external sources such as vital statistics.

Example 4.1. (Continued)
In order to obtain estimates of breast cancer risk in the presence of competing risks, Gail et al. (1989) used United States (US) mortality rates from year 1979 for all causes except breast cancer to estimate the competing risks with more precision than from the BCDDP follow-up data. In age groups 40–44 and 45–49 years, these death rates were 153.0 and 248.6 per 10^5 person-years, respectively, hence of the same order of magnitude as breast cancer incidence rates. In older age groups, these death rates were much higher than breast cancer incidence rates, thus strongly influencing breast cancer risk estimates for age intervals including these age groups. For instance, death rates from causes other than breast cancer were 1,017.7 and 2,419.8 per 10^5 person-years in age groups 65–69 and 70–74 years, respectively, whereas average incidence rates of breast cancer were 356.1 and 307.8 per 10^5 person-years in these age groups, respectively.

4.3.7 Estimation from Cohort Studies

Estimation of disease risk rests on estimating disease incidence and hazard rates, a topic also addressed in Part III (Statistical Methods in Epidemiology) of this handbook. Several approaches have been worked out fully for disease risk estimation. A brief review of these approaches is given here starting with average risk estimates that do not take exposure profiles into account and continuing with exposure-specific estimates.

4.3.7.1 Estimates of Average Disease Risk

The density or exponential method (Miettinen 1976; Kleinbaum et al. 1982, Chap. 6; Rothman and Greenland 1998, Chap. 3) relies on subdividing the time or age scale in successive time or age intervals $I_1,\ldots, I_i,\ldots, I_I$ (e.g., 1- or 5-year intervals) on which the rate of disease incidence is assumed constant (i.e., piecewise constant). Disease risk over time or age interval $[a_1, a_2)$, i.e., the probability for an individual to experience disease occurrence over interval $[a_1, a_2)$ is taken as one minus the probability of staying disease-free through the successive intervals included in $[a_1, a_2)$. Assuming that disease is rare on each of the successive intervals considered, disease risk can be estimated as

$$\hat{\pi}(a_1, a_2) = 1 - \exp\left(-\sum_i \hat{h}_i \Delta_i\right). \tag{4.5}$$

The sum is taken over all intervals included in $[a_1, a_2)$. Notation Δ_i denotes the width of interval i, whereas $\hat{h}_i$ denotes the incidence rate in interval i, obtained as the ratio of the number of incident cases over the person-time accumulated during follow-up in that interval.

While Eq. 4.5 is simple to apply, its validity depends on several assumptions. The assumption that disease incidence is constant over each time or age interval considered makes it a parametric approach. However, if intervals are small enough, this will not amount to a strong assumption. Moreover, it relies on the assumption that disease incidence is small on each interval. If this is not the case, a more complicated formula will be needed. Finally, this approach ignores competing risks.

Benichou and Gail (1990a) generalized this approach by lifting the condition on small incidence on each interval and allowing competing risks to be taken into account. They derived a generalized expression for the estimate of disease risk over time or age interval $[a_1, a_2)$ as

$$\hat{\pi}(a_1, a_2) = \sum_i \frac{\hat{h}_{1i}}{\hat{h}_{1i} + \hat{h}_{2i}} \left[1 - \exp\left\{-(\hat{h}_{1i} + \hat{h}_{2i})\Delta_i\right\}\right] A(i), \tag{4.6}$$

with $A(i) = \prod_{j<i} \exp\left\{-(\hat{h}_{1j} + \hat{h}_{2j})\Delta_j\right\}$.

In Eq. 4.6, the sum is taken over all intervals included in $[a_1, a_2)$, Δ_i denotes the width of interval i, $\hat{h}_{1i}$ denotes the disease incidence rate in interval i, $\hat{h}_{2i}$ the death rate from other causes in interval i, and the product in $A(i)$ is taken over time intervals in $[a_1, a_2)$ from the first one to the one just preceding interval i. Death rates can be obtained in a similar fashion as disease incidence rates. It should be noted that disease risk can be estimated for a much longer duration than the actual follow-up of individuals in the study if age is the time scale (open cohort) provided there is no secular trend in disease incidence.

Variance estimates were derived by Benichou and Gail (1990a). Moreover, based on simulations of a closed cohort, they found that resulting confidence intervals have satisfactory coverage, especially with the log transformation, and observed little or no bias on risk estimates with a sufficient number of intervals even when disease incidence varied sharply with time.

The actuarial method or life table method (Cutler and Ederer 1958; Elveback 1958; Fleiss et al. 1976; Kleinbaum et al. 1982, Chap. 6; Rothman and Greenland 1998, Chap. 3) shares similarities with the density method, although it was derived from a less parametric viewpoint. As with the density method, time is split into intervals. In each time interval i, the probability for an individual who is disease-free at the beginning of the interval to stay disease-free throughout the interval is estimated. Disease risk is obtained as one minus the estimated probability of staying disease-free throughout the successive time intervals included in $[a_1, a_2)$ as

$$\hat{\pi}(a_1, a_2) = 1 - \prod_i \frac{(n_i - w_i/2 - d_i)}{(n_i - w_i/2)}, \tag{4.7}$$

where the product is taken over all intervals included in $[a_1, a_2)$, n_i denotes the number of disease-free subjects at the beginning of interval i, d_i the number of incident cases occurring in interval i, and w_i the number of subjects either lost to follow-up or dying from other causes (competing risks) in interval i. The actuarial approach is most appropriate when grouped data are available and the actual follow-up of each individual in each interval is not known. The person-years of follow-up for subjects lost to follow-up or affected with competing risks in interval i is not used directly but, if one assumes that the mean withdrawal time occurs at the midpoint of the interval, then the denominator in each product term of Eq. 4.7 can be regarded as the effective number of persons at risk of developing the disease in the corresponding interval. Namely, it represents the number of disease-free persons that would be expected to produce d_i incident cases if all persons could be followed for the entire interval (Elandt-Johnson 1977; Kleinbaum et al. 1982, Chap. 6; Littell 1952). The actuarial method can be regarded as a refinement of the simple cumulative method (Kleinbaum et al. 1982, Chap. 6) that ignores quantity w_i and simply estimates disease risk as the number of individuals who contract the disease divided by the total number in the cohort or exposure subgroup of interest. The actuarial method is preferable to this direct method because, in practice, it is rare that a large enough cohort can be followed over a long enough time to reliably estimate the risk of disease by this simple method. Moreover, the simple cumulative method cannot handle the case when subjects are followed for varying lengths of time, which often occurs because subjects can be enrolled at different times whereas the follow-up ends at the same time for all subjects.

As shown by several authors (Cutler and Ederer 1958; Fleiss et al. 1976), the actuarial method results in biased estimates of risk even in the unlikely and most favorable event (in terms of bias) of all withdrawals occurring at the interval midpoints. Alternative approaches based on different choices of the quantity to subtract from n_i (i.e., choices different from $w_i/2$) are not subject to less bias, however (Elandt-Johnson 1977). The problem can be best handled by using narrow intervals, but this is done at the expense of a larger random error (i.e., less precise estimates of risk).

Compared to the density method (Eqs. 4.5 and 4.6), the actuarial method has the advantage of not requiring knowledge of individual follow-up times in each interval but only knowledge of the number at risk at the beginning of the interval and the number of withdrawals. The density method could be used, however, without knowledge of follow-up time by assigning a follow-up time of half the interval width to subjects who are lost to follow-up, develop disease or die from other causes, in an analogous fashion as with the actuarial method (Benichou and Gail 1990a). The actuarial method requires neither the assumption of constant incidence rate nor rarity of disease incidence on all time intervals. However, bias is less of a problem with the density than the actuarial method and the density method applies naturally to open cohorts and extends easily to risk estimates that take exposure profiles into account (see below).

When individual follow-up times are all known, a fully non-parametric risk estimate can be obtained in the spirit of the Kaplan–Meier estimate of survival (Kaplan and Meier 1958; see also chapter ▸Survival Analysis of this handbook). Disease risk is estimated through summation on all distinct times in $[a_1, a_2)$ at which new disease cases occur (Aalen and Johansen 1978; Kay and Schumacher 1983; Gray 1988; Matthews 1988; Keiding and Andersen 1989; Benichou and Gail 1990a; Korn and Dorey 1992). Corresponding variance estimates were derived (Aalen 1978; Aalen and Johansen 1978; Keiding and Andersen 1989; Benichou and Gail 1990a; Korn and Dorey 1992) from which confidence intervals can be obtained, based on the log transformation as suggested by Benichou and Gail (1990a) and Keiding and Andersen (1989) or based on the approach of Dorey and Korn (1987).

Upon comparing the generalized density method (see Eq. 4.6) and the non-parametric method, Benichou and Gail (1990a) showed that the loss of efficiency of the non-parametric method is small compared to the density method. Moreover, the non-parametric method yields little bias in risk estimates as well nearly nominal coverage for confidence intervals of risk with the log transformation. Nominal coverage refers to the theoretical probability of a confidence interval to cover the true parameter and may be assessed using simulations (i.e., a 95% confidence interval will be said to have nominal coverage if it does include the true parameter value in 95% of the cases). Hence, properties of the generalized density and non-parametric methods agree closely. However, the generalized density method has the advantage of simplicity of computation and is better suited to open cohorts.

4.3.7.2 Estimates of Exposure-Specific Disease Risk

In order to obtain risk estimates that depend on exposure profiles, the cohort could be subdivided into subcohorts based on exposure levels and the methods above applied to these subcohorts. However, this approach would be impractical because it would yield risk estimates with very low precision. In order to remedy this problem, a natural approach to incorporate exposures is to model incidence rates through regression models.

Benichou and Gail (1990a) proposed a direct extension of the generalized density method (Eq. 4.6). This extension is based on assuming that the disease hazard rate on each time or age interval i is the product of a constant baseline hazard rate for subjects at the reference level of exposure in interval i and a function of the various exposures. The corresponding parameters, i.e., baseline hazard rates and hazard ratio parameters for exposure, can be jointly estimated by maximizing the piecewise exponential likelihood, which is equivalent to the usual Poisson likelihood for the analysis of cohort data (Holford 1980; Laird and Oliver 1981). Corresponding variance estimates are available (Benichou and Gail 1990a). In simulations, risk estimates appeared subject to little bias, variance estimates were also little biased, and coverage of confidence intervals was nearly nominal, except for the exposure profiles with very few subjects (Benichou and Gail 1990a). Other parametric approaches were considered to obtain risk estimates of cardiovascular events from the Framingham studies (Anderson et al. 1991). Semi-parametric estimators of risk were also derived (Benichou and Gail 1990a). In contrast with the previous approach where a piecewise exponential or Poisson distribution is assumed, the baseline disease hazard rate is expressed as an unspecified function of time or age rather than a constant, which corresponds to the semi-parametric Cox regression model (Cox 1972). Risk estimates are obtained as functions of the partial likelihood estimates (Cox 1975) of hazard ratio parameters and related Nelson-Aalen estimates of cumulative baseline hazard rates (Borgan 1998). From results in Tsiatis (1981) and Andersen and Gill (1982) on the joint distribution of these parameter estimates, Benichou and Gail (1990a) derived an asymptotic variance estimator.

Regression based methods appear well suited for estimating exposure-specific disease risk and are therefore useful for the purpose of individual prediction. Compared to the semi-parametric approach, the generalized density method appears easier to implement while providing a good compromise between bias and precision.

4.3.8 Estimation from Population-Based or Nested Case-Control Studies

As discussed above, whereas disease risk is directly estimable from cohort data, case-control data have to be complemented with follow-up or population data in order to obtain the necessary information on incidence rates. If such complementary data are available, exposure-specific incidence rates and exposure-specific disease risk can be estimated. All approaches proposed in the literature rely on regression methods.

4.3.8.1 The Hybrid Approach

This approach relies on the assumption of piecewise constant incidence rates and on Eq. 4.2 to obtain baseline incidence rates in strata defined by factors such as age, sex, race, or geographical area (see Sect. 4.2.2). Odds ratio estimates are then combined with baseline incidence rates to arrive at exposure-specific incidence rates (see Sect. 4.2.2). Applying Eq. 4.6 to these rates and death rates from competing causes, disease risk estimates can be obtained for desired time intervals. This approach has been used in practice to obtain individual risks of breast cancer by Gail et al. (1989) (see Example 4.1 below). Resulting disease risk estimates can be termed estimates of individual breast cancer risk since they depend on age and individual exposure profile (216 profiles were considered overall). The approach can be seen as a multivariate extension of earlier work by Miettinen (1974). It has been termed a hybrid approach (Benichou 2000a) since it relies on two models, namely, the piecewise exponential model that underlies the density method (i.e., constant incidence by age group) and the logistic model used to obtain odds ratio estimates from the nested case-control data (see Sect. 4.4.6.3). It can be applied to population-based case-control data with no individual follow-up of subjects in a similar manner as to nested case-control data, as discussed and illustrated for bladder cancer by Benichou and Wacholder (1994) (see Example 4.2 below).

Variance estimators for risk estimates are complex since exposure-specific incidence rate estimates involve odds ratio parameters obtained through logistic regression from the case-control data and counts of incident cases from the follow-up or population data. Estimators of variances and covariances of age- and exposure-specific incidence rates that take into account all sources of variability have been fully worked out for various sampling schemes regarding control selection in the general case (Benichou and Gail 1990a) and specifically to account for the special features of the BCDDP data (Benichou and Gail 1995). Simulations tailored to the BCDDP data showed a small upward bias in risk estimates due to the small upward bias incurred by using odds ratios to estimate hazard ratios when the rare-disease assumption appeared questionable. Variance estimates had very little bias and yielded confidence intervals with near nominal coverage. Coverage was improved with the logit transformation.

Example 4.1. (Continued)

Applying Eq. 4.6 to exposure-specific incidence rates of breast cancer estimated from the BCDDP data (see Sect. 4.2.2) and death rates from other causes estimated from US mortality data (see Sect. 4.3.6), risk estimates of breast cancer can be obtained. For instance, the 10-year risk of developing breast cancer between ages 40 and 50 years for a woman initially free of breast cancer at age 40 years and with the exposure profile considered in Sect. 4.2.2 (i.e., nulliparous woman with menarche at age 12 years, one previous biopsy for benign breast disease, and no history of breast cancer in her first-degree relatives) is obtained as a sum of 2 terms. The first term, $\hat{\pi}_1$, corresponding to age interval 40–44, is obtained from Eq. 4.6 as

$$\begin{aligned}\hat{\pi}_1 = &\frac{307.8 \times 10^{-5}}{307.8 \times 10^{-5} + 153.0 \times 10^{-5}}\\ &\times \left[1 - \exp\left\{-5(307.8 \times 10^{-5} + 153.0 \times 10^{-5})\right\}\right] = 0.0152.\end{aligned}$$

The second term, $\hat{\pi}_2$, corresponding to age interval 45–49, is obtained from Eq. 4.6 as the product of the probability of developing breast cancer in age interval 45–49 times the probability of having stayed free of breast cancer and not died from other causes in age interval 40–44:

$$\begin{aligned}\hat{\pi}_2 = & \frac{424.3 \times 10^{-5}}{424.3 \times 10^{-5} + 248.6 \times 10^{-5}} \\ & \times \left[1 - \exp\left\{-5(424.3 \times 10^{-5} + 248.6 \times 10^{-5})\right\}\right] \\ & \times \exp\left\{-5(307.8 \times 10^{-5} + 153.0 \times 10^{-5})\right\} = 0.0204.\end{aligned}$$

Thus, the 10-year risk of developing breast cancer is obtained as the sum $0.0152+0.0204 = 0.0356$, or 3.6%. The corresponding 95% confidence interval based on taking all sources of variability into account can be estimated as 3.0–4.2% through computations described in Benichou and Gail (1995). Breast cancer risk estimates can be obtained for all age intervals in the range 20–80 years and all 216 exposure profiles including the profile considered above. This whole approach to individual breast cancer risk estimation is known as the "Gail model" and has enjoyed widespread use in individual counseling, and in designing, and interpreting prevention trials. Practical implementation has been greatly facilitated by the development of graphs (Benichou et al. 1996) as well as a computer program (Benichou 1993a) and its modified version that is available on the US National Cancer Institute web site at http://bcra.nci.nih.gov/brc/ accessed 12 May 2004.

Example 4.2.
In the year 1978, incident cases of bladder cancer were identified through ten cancer registries in the Unites States. For instance, 32 incident cases were identified among white males aged 45–64 years whose population numbered 97,420 individuals. Assuming that this population remained constant throughout the year 1978, these data yielded an average incidence rate of 32.8 per 10^5 person-years. The National Bladder Cancer Study was a population-based case-control study conducted at the ten cancer registries. Incident cases aged 21–84 years were selected from the registries. Controls aged 21–84 years were selected from telephone sampling or Health Care Financing Administration rosters and frequency-matched to cases on geographical area, age, and sex. Based on case-control data from two states (Utah and New Jersey) and one large city (Atlanta), odds ratios were estimated for smoking status (never smoker, ex-smoker, current light smoker, current heavy smoker) and occupational exposure to carcinogens (yes, no) using logistic regression (see Sect. 4.4.6.3). Moreover, the attributable risk for smoking and occupational exposure was estimated for white males in each of the nine strata resulting from the three areas and three age groups (i.e., 21–44, 45–64 and 65+ years) (see Sect. 4.5.1). Among white males aged 45–64 years in Utah, it was estimated at 54.0%, yielding a baseline incidence rate of $32.8 \times (1 - 0.540) = 15.1$ per 10^5 person-years. The odds ratios for current heavy smokers ($\geq$20 cigarettes per day) and occupational exposure were estimated at 2.9 and 1.6. Hence, among white males aged 45–64 years in Utah, exposure-specific incidence rates were estimated at $15.1 \times 1.6 = 24.1$ per 10^5 person-years for never smokers with a history of occupational exposure and $15.1 \times 2.9 \times 1.6 = 69.8$ per 10^5 person-years for current heavy smokers with a history of occupational exposure assuming a multiplicative effect of smoking and occupational exposure (and allowing for rounding error). From these exposure-specific incidence rates, estimates of the risk of bladder cancer over specified age intervals could be derived, using Eq. 4.6.

4.3.8.2 Other Parametric Approaches

A pseudo-likelihood approach also relying on the assumption on piecewise constant incidence (i.e., piecewise exponential model) has been proposed as an alternative to the hybrid approach (Benichou and Wacholder 1994). In each stratum separately, observed distributions of exposure in the cases and controls are applied to counts of incident cases and person-time to obtain respective expected numbers of incident cases and of person-time per stratum and exposure level. Then, baseline incidence rates and hazard ratios are jointly estimated from these expected quantities under a piecewise exponential model. Joint estimation proceeds from maximizing the likelihood corresponding to this model. Since this likelihood includes expected rather than observed counts, it is termed a pseudo-likelihood. Thus, the procedure includes two steps. In the first step, expected numbers of incident cases and person-time per exposure and stratum are calculated. Then, the parameters of interest (i.e., stratum-specific baseline incidence rates and hazard ratios) are estimated from these expected counts through maximizing a pseudo-likelihood. This approach is easy to implement, as was illustrated on population-based case-control data of bladder cancer.

Example 4.2. (Continued)
Among white males aged 45–64 years and in all other strata separately, observed proportions of cases (respectively controls) with given joint level of smoking and occupational exposure among the eight (four times two) joint levels considered were applied to counts of incident cases (respectively person-time) to obtain expected counts by stratum and joint exposure level. Namely, the products of the counts by the observed proportions were formed. Using these expected counts, a pseudo-likelihood based on the piecewise exponential model was maximized yielding estimates of relative hazards and stratum-specific baseline incidence rates. For instance, the baseline incidence rate for white males aged 45–64 years in Utah was estimated at 13.7 per 10^5 person-years, and relative hazards for current heavy smoking and occupational exposure were estimated at 2.9 and 1.5, respectively. Hence, among white males aged 45–64 years in Utah, exposure-specific incidence rates were estimated at $13.7 \times 1.5 = 20.6$ per 10^5 person-years for never smokers with a history of occupational exposure and $13.7 \times 2.9 \times 1.5 = 61.9$ per 10^5 person-years for current heavy smokers with a history of occupational exposure, still assuming a multiplicative effect of smoking and occupational exposure (and allowing for rounding error).

A full likelihood approach has also been proposed based on the piecewise exponential model (Benichou and Wacholder 1994). All parameters (i.e., baseline rates, hazard ratios, and conditional probabilities for the distribution of exposure in the cases and controls) are estimated jointly through maximizing a likelihood involving all parameters. This approach may prove intractable in practice except in simple situations with few exposure levels considered. A full likelihood approach based on the logistic model (Greenland 1981) appears much easier to implement. Baseline incidence rates are obtained by simply adding to the stratum parameter estimates from the logistic model a term corresponding to the logarithm of the ratio of sampling fractions among cases and controls in the stratum (Greenland 1981; Prentice and Pyke 1979; also similar to discussion of Eq. 4.11 in Sect. 4.4.6.3).

Example 4.2. (Continued)
Although it required the estimation of 60 additional parameters relative to the pseudo-likelihood approach, the full likelihood approach based on the piecewise exponential model could be implemented. The 60 additional parameters described the conditional probabilities of exposure (smoking and occupational exposure) in the cases and controls for all nine strata. For instance, the baseline incidence rate for white males aged 45–64 years in Utah

was estimated at 13.9 per 10^5 person-years, and relative hazards for current heavy smoking and occupational exposure were estimated at 2.9 and 1.6, respectively. Hence, among white males aged 45–64 years in Utah, exposure-specific incidence rates were estimated at $13.9 \times 1.6 = 22.2$ per 10^5 person-years for never smokers with a history of occupational exposure and $13.9 \times 2.9 \times 1.6 = 64.1$ per 10^5 person-years for current heavy smokers with a history of occupational exposure still assuming a multiplicative effect of smoking and occupational exposure (and allowing for rounding error).

Upon comparing the pseudo-likelihood, full likelihood and hybrid approach on population-based case-control data of bladder cancer, Benichou and Wacholder (1994) noted that the hybrid approach seemed to be less efficient for incidence rate estimation than the other two approaches, which were themselves equally efficient. They discussed other advantages of the pseudo-likelihood and full likelihood approaches. Namely, these approaches allow direct estimation of hazard ratios rather than odds ratios. Furthermore, the pseudo-likelihood approach and the full likelihood approach (in its version relying on the piecewise exponential model) can be applied to more general regression models, e.g., models with an additive form using hazard rate difference parameters rather than hazard ratio parameters (see Sects. 4.4.4 and 4.4.6.4). Finally, all three approaches require that cases and controls be selected completely at random and that incident cases or at least a known proportion of them (i.e., known sampling fraction) be fully identified.

4.3.8.3 Semi-Parametric Approach

In nested case-control studies, controls are usually individually matched to cases on time. Namely, for each case, one (or several) control(s) is (are) selected among subjects with the same age and length of follow-up in the cohort as the case (Breslow et al. 1983; Liddell et al. 1977; Mantel 1973; see also chapter ▸Modern Epidemiological Study Designs of this handbook).

The three parametric approaches described in Sects. 4.3.8.1 and 4.3.8.2 do not apply readily to this context of individual time matching of controls to cases. Langholz and Borgan (1997) developed a semi-parametric approach to handle this case. Their approach can be regarded as an extension of the semi-parametric approach for cohort studies described above (see Sect. 4.3.7.2). Incidence rates are expressed as the product of baseline incidence rates of an unspecified form times a function of the covariates representing the hazard ratio (Cox 1972). Hazard ratio parameter estimates are obtained from maximizing the partial likelihood of the Cox model for nested case-control data (Oakes 1981; Prentice and Breslow 1978). Risk estimates are obtained by combining partial likelihood hazard ratio parameter estimates and corresponding cumulative hazard estimates.

A direct comparison of the semi-parametric approach with the parametric approaches presented in Sects. 4.3.8.1 and 4.3.8.2 is not possible because the semi-parametric approach applies only to time-matched data, which the parametric approaches cannot handle. The semi-parametric approach requires observation of individual follow-up time of each subject in the original cohort in order to form the risk sets for each failure time and enable control selection. It is therefore potentially less widely applicable than the parametric approaches.

4.3.9 Final Notes and Additional References

General problems of definition of disease risk, interpretation and usefulness, properties, estimation, and special problems have been reviewed in detail (Benichou 2000a). Special problems include accounting for continuous or time-dependent exposure, estimation of disease risk from two-stage case-control data, and validation procedures for disease risk estimates. Finally, an important challenge is to increase awareness of the proper interpretation and use of disease risk in practice and develop general software for easier implementation.

4.4 Measures of Association

4.4.1 Definitions and General Points

Measures of association have a long history and have been reviewed in many textbooks. They assess the strength of associations between one or several exposures and the risk of developing a given disease. Thus, they are useful in etiological research to assess and quantify associations between potential risk (or protective) factors and disease risk. The question addressed is whether and to what degree a given exposure is associated with occurrence of the disease of interest. In fact, this is the primary question that most epidemiological studies are trying to answer.

Depending on the available data, measures of association may be based on disease rates, disease risks, or even disease odds, i.e., $\pi/(1-\pi)$, with π denoting disease risk. They contrast rates, risks, or odds for subjects with various levels of exposure, e.g., risks or rates of developing breast cancer for 40-year-old women with or without a personal history of benign breast disease. They can be expressed in terms of ratios or differences of risks or rates among subjects exposed and non-exposed to given factors or among subjects with various levels of exposure.

Measures of association can be defined for categorical or continuous exposures. For categorical exposures, any two exposure levels can be contrasted using the measures of association defined below. However, it is convenient to define a reference level to which any exposure level can be contrasted. This choice is sometimes natural (e.g., non-smokers in assessing the association of smoking with disease occurrence) but can be more problematic if the exposure considered is of continuous nature, where a range of low exposures may be considered potentially inconsequential. The choice of a reference range is important for interpreting results. It should be wide enough for estimates of measures of association to be reasonably precise. However, it should not be so wide that it compromises meaningful interpretation of the results, which depend critically on the homogeneity of the reference level. For continuous exposures, measures of association can also be expressed per unit of exposure, e.g., for each additional gram of daily alcohol consumption. The reference level may then be a precise value such as no daily alcohol consumption or a range of values such as less than 10 g of daily alcohol consumption.

4.4.2 Usefulness and Interpretation

When computing a measure of association, it is usually assumed that the relationship being captured has the potential to be causal and efforts are taken to remove the impact of confounders from the quantity. Section 4.4.6 provides a summary of techniques for adjustment for confounders. Nonetheless, except for the special case of randomized studies, most investigators retain the word "association" rather than "effect" when describing the relationship between exposure and outcome to emphasize the possibility that unknown confounders may still influence the relationship.

Rothman and Greenland (chapter ▶Basic Concepts of this handbook) take efforts to differentiate the concepts of effect and association, and adopt the framework of *counterfactuals*, popular in the field of economics (Wooldridge 2001), to define the term *effect size*. They then define "measure of association" as computed to compare two actual populations. Hence, the distinction is one of a true causal concept versus one that may be subject to the confounding of the true effect arising from the population mix of characteristics at hand. These definitions are more precise and serve as reminders of the true nature of causality. We will retain the less precise but more common terminology where "measure of association" refers to either or both concepts. We also note that the discussion here is limited to measures of association with a *binary* (i.e., coded as 1 = present, 0 = absent) or event *count* (number of events) outcome. In many situations, classification into disease versus no disease is not clear-cut. For example, the definition of an abnormal lipid profile has undergone frequent change. In such cases, using measures based on continuous outcomes may be a better choice. We comment on relationships between measures of association for continuous and categorical outcomes in Sect. 4.4.6.2.

When choosing a measure of association, the primary goal is interpretability and familiarity to consumers of the information. Another guideline is that the measure of association should allow as simple a description of the association as possible. For example, it has been empirically observed that risk ratios are more likely than risk differences to remain constant across subpopulations with different risk levels (Breslow and Day 1980, Chap. 2), hence simplifying description of the association of the exposure with the outcome. Breslow and Day (1980, Chap. 2), also point out that ratios can be converted to differences by taking the logarithm of the risk or rate.

Definitions and properties of measures of association as well as relations among them are reviewed below for measures based on ratios and measures based on differences. Then, estimability of these measures from cohort and case-control designs and general points regarding estimation of these measures are considered, including an overview of techniques to adjust for confounders. More details regarding inference, namely, estimating these measures and assessing the statistical

Table 4.1 Measures of association discussed in this chapter (GLM = generalized linear model; see Sect. 4.4.6)

Measure	Lower limit	Upper limit	Null value	Definition	Link function in GLM
Rate ratio (*HR*)	0	$+\infty$	1	$h_E/h_{\bar{E}}$	Log
Risk ratio (*RR*)	0	$+\infty$	1	$\pi_E/\pi_{\bar{E}}$	Log
Odds ratio (*OR*)	0	$+\infty$	1	$[\pi_E/(1\text{-}\ \pi_E)]/[\pi_{\bar{E}}\ /(1\text{-}\pi_{\bar{E}})]$	Logit
Rate difference	$-\infty$	$+\infty$	0	$h_E - h_{\bar{E}}$	Identity
Risk difference	-1	$+1$	0	$\pi_E\text{-}\ \pi_{\bar{E}}$	Identity

significance of apparent associations, will be presented in Part III (Statistical Methods in Epidemiology) of this handbook.

The above Table 4.1 provides an overview of measures of association discussed in this chapter.

4.4.3 Measures Based on Ratios

4.4.3.1 General Properties

Ratio-based measures of association are particularly appropriate when the effect of the exposure is multiplicative, which means there is a similar percent increase or decrease associated with exposure in rate, risk, or odds across exposure subgroups. As noted above, effects have often been observed to be multiplicative, leading to ratios providing a simple description of the association (e.g., see Breslow and Day 1980, Chap. 2). Ratio measures are dimensionless and range from zero to infinity, with one designating no association of the exposure with the outcome. When the outcome is death or disease, and the ratio has the rate, risk or odds of the outcome with the exposed group in the numerator, a value less than one indicates a protective effect of exposure. The exposure is then referred to as a protective factor. When the ratio in this setup is greater than one, there is greater disease occurrence with exposure, and the exposure is then referred to as a risk factor.

It can be shown that numerically, the odds ratio falls the furthest from the null, and the risk ratio the closest, with the rate ratio in between. For example, from the below Table 4.2, based on a fictitious data from a cohort study for a disease that is not rare, we would obtain a risk ratio $\widehat{RR} = 0.3/0.1 = 3.00$ and an odds ratio $\widehat{OR} = [(30)(90)]/[(10)(70)] = 3.86$. If we assume a constant hazard rate, so that the risk for each group is $1 - \exp(-hT)$, with T being the follow-up time for each subject, we have the rate ratio $\widehat{HR} = \ln(1-0.3)/\ln(1-0.1) = 3.39$ (calculations described in Sect. 4.4.6.1). Hence, $1 < \widehat{RR} < \widehat{HR} < \widehat{OR}$.

The difference in magnitude between the above ratio measures is important to keep in mind when interpreting them for diseases or outcomes that are not rare. For rare outcomes, the values of the three ratio measures tend to be close. Ratios become differences on the logarithmic scale, and estimation and inference often take place on the log scale, where zero indicates no association.

Table 4.2 Data from fictitious cohort study

	Exposed	Unexposed
Diseased	30	10
Non-diseased	70	90

4.4.3.2 Rate Ratios

As the name implies, the rate ratio is the ratio between the rate of disease among those exposed and those not exposed or $h_E/h_{\bar{E}}$. Conceptually, the rate ratio is identical to a hazard ratio HR. The latter term tends to be used when time dependence of the rate is emphasized, as the hazard rate is a function that may depend on time. The situation of a constant rate ratio over time is referred to as *proportional hazards*. The proportional hazard assumption is often made in the analysis of rates (see below). Theoretically, the hazard ratio at a given time point is the limiting value of the rate ratio as the time interval around the point becomes very short, just as the hazard rate is the limiting quantity for incidence rate (see Sect. 4.2). The rate ratio has also been called the *Incidence Density Ratio* (Kleinbaum et al. 1982, Chap. 8). It may be noted that the rate ratio is attenuated by less than perfect specificity of the outcome criteria, but relatively unaffected by less than perfect sensitivity, especially when the rate is low, as long as the sensitivity is unaffected by exposure. In other words, if cases are equally missed in the exposed and unexposed groups, the rate ratio is relatively unaffected. However, if non-cases are considered cases, the ratio will be lower than if diagnostic criteria identified only true cases. Even in the fictitious example above with high incidence rates, 80% sensitivity leads to a slightly attenuated rate ratio of 3.29 from

$$\widehat{HR} = \ln[1 - (0.80)(0.3)]/\ln[1 - (0.80)(0.1)) = 3.29$$

(as compared to the correct rate ratio of 3.39 from Table 4.2), while 80% specificity leads to a severely biased rate ratio of

$$\widehat{HR} = \ln[0.80(1 - 0.3)]/\ln[0.80(1 - 0.1)] = 1.77.$$

Rate ratios are extremely useful because of the ease of estimating them in many contexts. They refer to population dynamics, and are not as easily interpretable on the individual level. It has been argued, however, that rate ratios make more sense than risk ratios (see Sect. 4.4.3.3) when the period subjects are at risk is longer than the observation period (Kleinbaum et al. 1982, Chap. 8). Numerically, the rate ratio is further from the null than the risk ratio. When rates are low, the similarity of risk and rate leads to rate ratios being close to risk ratios, as discussed in Sect. 4.4.3.3. Some investigators tend to refer to rate ratios as relative risks, creating some confusion in terminology. Further considerations of how the rate ratio relates to other ratio-based measures of association are offered by Rothman and Greenland (1998, p. 50).

4.4.3.3 Risk Ratios

The risk ratio, relative risk, or ratio of risks of disease among those exposed π_E and those not exposed $\pi_{\bar{E}}$, $RR = \pi_E/\pi_{\bar{E}}$, has been viewed as the gold standard among measures of association for many years. It is eminently interpretable on the individual level as a given-fold increase in risk of disease. Like other ratio-based measures, it tends to be more stable than the risk difference across population groups at widely different risk. However, similar to rate ratios and odds ratios (introduced in Sect. 4.4.3.4), the risk ratio can be viewed as misleading in the public eye when the risk among both the unexposed and the exposed is very low, yet many-fold increased by exposure. Another disadvantage of the risk ratio is its asymmetry with respect to the definition of an event, so that the risk ratio for not having an event, $(1 - \pi_E)/(1 - \pi_{\bar{E}})$, cannot be directly computed from the risk ratio for having an event. For example, knowing that the risk ratio for an event $RR = 3.00$, the scenario $\pi_E = 0.3$, $\pi_{\bar{E}} = 0.1$ results in $(1 - \pi_E)/(1 - \pi_{\bar{E}}) = 0.7/0.9 = 0.78$, while the scenario $\pi_E = 0.6$, $\pi_{\bar{E}} = 0.2$, which represents the same risk ratio of 3.00, results in $(1-\pi_E)/(1-\pi_{\bar{E}}) = 0.4/0.8 = 0.50$. The risk ratio depends on the length of the time interval considered because risk itself refers to a specific interval (see Sect. 4.3). In the literature, the term relative risk is often used to denote the rate ratio as well as the risk ratio, creating some confusion. Therefore, we will avoid the term "relative risk" in the following. Numerically, the risk ratio is closer to the null than the rate ratio for the same data (see Sect. 4.4.3.1).

Cornfield et al. (1959), in the smoking versus lung cancer debate, derived several theoretical properties of the risk ratio, which have further supported its use. In this debate, Cornfield, along with Doll and Hill, argued against strong opposition from R.A. Fisher and Joseph Berkson that the association was causal and not likely due to unmeasured confounders, such as a genetic predisposition to both smoke and contract lung cancer. First of all, Cornfield et al. (1959) turned attenuation of the risk ratio due to lack of specificity of the outcome into an advantage, by noting that the ratio will become stronger as the disease subtype affected by the exposure is honed. Second, Cornfield et al. demonstrated that if a confounder is to explain the outcome with exposure risk ratio $RR > 1$, that confounder has to have risk ratio at least *RR*, and in addition, the prevalence of the confounder must be at least *RR* times greater among the exposed than among the unexposed. Lin et al. (1998) presented more general formulas that confirm Cornfield et al.'s assertions under assumptions of no interaction between the confounder and exposure. These theoretical results have led investigators to reason that high risk ratios (say above 1.4; Siemiatycki et al. 1988) are not likely to be explained by uncontrolled confounding.

4.4.3.4 Odds Ratios

For several reasons, the odds ratio has emerged as the most popular measure of association. The odds ratio is the ratio of odds, $OR = [\pi_E/(1 - \pi_E)]/[\pi_{\bar{E}}/(1 - \pi_{\bar{E}})]$. Historically, the odds ratio was considered an approximation to the risk ratio obtainable from case-control studies. The reason for this is that the probabilities of being sampled into case and control groups cancel in the calculation of the odds ratio, as long as sampling is independent of exposure status. Furthermore, when π_E and $\pi_{\bar{E}}$ are small, the ratio $(1 - \pi_{\bar{E}})/(1 - \pi_E)$ has little influence on the odds ratio, making it approximately equal to the risk ratio $\pi_E/\pi_{\bar{E}}$. The assumption of small π_E and $\pi_{\bar{E}}$ is referred to as the *rare-disease assumption.* Kleinbaum et al. (1982) have pointed out that in a case-control study of a stable population with incident cases and controls being representative of non-cases, the odds ratio is the rate ratio. Numerically, the odds ratio is the furthest from the null of the three ratio measures considered here.

More recently, the odds ratio has gained status as an association measure in its own right and is often applied in cohort studies and clinical trials, as well as in case-control studies. This is due to many desirable properties of the odds ratio. First of all, focusing on risk rather than odds may be a matter of convention rather than a preference based on fundamental principles, and using the same measure across settings has the advantage of consistency and makes comparisons and meta-analyses easy. In contrast to the risk ratio, the odds ratio is symmetric so that the odds ratio for disease is the inverse of the odds ratio for no disease. Furthermore, the odds ratio based on exposure probabilities equals the odds ratio based on disease probabilities, a fact that follows from Bayes' theorem (e.g., Cornfield 1951; Miettinen 1974; Neutra and Drolette 1978) or directly from consideration of how cases and controls are sampled. The disease and exposure odds ratios are sometimes referred to as *prospective* and *retrospective odds ratios*, respectively. Finally, odds ratios from both case-control and cohort studies are estimable by logistic regression, which has become the most popular approach to regression analysis with binary outcomes (see Sect. 4.4.6.3).

Some investigators feel that the risk ratio is more directly interpretable than the odds ratio and have developed methods for converting odds ratios into risk ratios for situations when risks are not low (Zhang and Yu 1998).

One disadvantage of odds ratios is that the crude odds ratio differs from an odds ratio conditional on (adjusted for) another factor that is related to the outcome, even when that factor is unrelated to the exposure, i.e. not a confounder (Miettinen and Cook 1981). This problem, referred to as non-collapsibility does not occur for risk ratios and risk differences. Non-collapsibility is related to the modeling considerations of cluster specific and marginal logistic models, and to the well-known difference between matched and unmatched odds ratio estimates, both discussed in Sect. 4.4.6.3.

4.4.4 Measures Based on Differences

4.4.4.1 General Properties

Difference-based measures lead to simple models when effects are additive (e.g., see Breslow and Day 1980, Chap. 2), which means that the exposure leads to a similar absolute increase or decrease in rate or risk across subgroups. The difference in odds is very rarely used and not addressed here. As noted above, additive relationships are less common in practice, except on the logarithmic scale, when they are equivalent to ratio measures. However, difference measures may be more understandable to the public when the outcome is rare, and relate directly to measures of impact discussed below (see Sect. 4.5).

The numerical ranges of difference measures depend on their component parts. The rate difference ranges from minus to plus infinity, while the risk difference is bounded between minus and plus one. The situation of no association is reflected by a difference measure of zero. When the measure is formed as the rate or risk among the exposed minus that among the non-exposed, a positive value indicates that the exposure is a risk factor, while a negative value indicates that it is a protective factor. It can be shown that the risk difference over a given time period falls numerically nearer to the null than does the rate difference expressed per the same period as the unit. For example, Table 4.2 yields a risk difference of $0.30 - 0.10 = 0.20$, while the rate difference is $-\ln(0.70) + \ln(0.90) = 0.25$. However, they will be close for rare outcomes. In contrast to ratio measures, difference measures are always attenuated by less than perfect sensitivity (i.e., missed cases), but the rate difference is unaffected by less than perfect specificity. The risk difference is also relatively unaffected when risk is low. In the fictitious example above, if the sensitivity of the test used to detect disease is 80%, the rate difference is $-\ln[1 - (0.80)(0.3)] + \ln[1 - (0.80)(0.1)] = 0.19$, but if the specificity is 80%, the rate difference remains at 0.25.

4.4.4.2 Rate Differences

The rate difference is defined as $h_E - h_{\bar{E}}$ and has been commonly employed to compare mortality rates and other demographic rates between countries, time periods, and/or regions. In such comparisons, the two rates being compared are often *directly standardized* (see Sect. 4.6) to the age and sex distribution of a standard population chosen, e.g., as the population of a given country in a given census year.

For the special case of a dichotomous exposure, the rate difference, i.e., the difference between the incidence rates in the exposed and unexposed subjects has been termed “excess incidence” (Berkson 1958; MacMahon and Pugh 1970; Mausner and Bahn 1974), “excess risk” (Schlesselman 1982), “Berkson’s simple difference” (Walter 1976), “incidence density difference” (Miettinen 1976), or even “attributable risk” (Markush 1977; Schlesselman 1982), which may have caused some confusion.

4.4.4.3 Risk Differences

The risk difference $\pi_E - \pi_{\bar{E}}$ is parallel to the rate difference discussed in Sect. 4.4.4.2, and similar considerations apply. Due to the upper and lower limits of plus, minus one on risk, but not on rate, risk differences are more difficult to model than rate differences.

4.4.5 Estimability

Because exposure-specific incidence rates and risks can be obtained from cohort data, all measures of association considered (based on ratios or differences) can be obtained as well. This is also true of case-control data complemented by follow-up or population data (see Sects. 4.2 and 4.3). Case-control data alone allow estimation of odds ratios thanks to the identity between disease and exposure odds ratios (see Sect. 4.4.3.4) that extends to the logistic regression framework. Prentice and Pyke 1979 showed that the unconditional logistic model (see also Breslow and Day 1980, Chap. 6) applies to case-control data as long as the intercept is disregarded (see Sect. 4.4.6.3). Interestingly, time-matched case-control studies allow estimation of hazard rates (e.g., see Miettinen 1976; Greenland and Thomas 1982; Prentice and Breslow 1978).

4.4.6 Estimation

The most popular measures of association have a long history of methods for estimation and statistical inference. Some traditional approaches have the advantage of being applicable in small samples. Traditional methods adjust for confounders by *direct standardization* (see Sect. 4.6.1) of the rates or risks involved, prior to computation of the measure of association, or by *stratification*, where association measures are computed separately for subgroups and then combined. For measures based on the difference of rates or risks, direct standardization and stratification can be identical, if the same weights are chosen (Kahn and Sempos 1989). Generally, however, direct standardization uses predetermined weights chosen for external validity, while *optimal* or *efficient* weights are chosen with stratification. Efficient weights make the standard error of the combined estimator as small as possible. *Regression adjustment* is a form of stratification, which provides more flexibility, but most often relies on large sample size for inference.

In modern epidemiology, measures of association are most often estimated from regression analysis. Such methods tend to require large sample sizes, in particular when based on *generalized linear models* (often abbreviated *GLM*). In this context, the ratio, difference, or other association measures arise from the regression coefficient of the exposure indicator, and different measures of association result depending on the transformation applied to the mean of the outcome variable. Note that the mean of an event count over a unit time interval is the rate, and the mean of a binary outcome is the risk. For example, a model may use the logarithm of the rate ($\ln(h)$) or risk ($\ln(\pi)$) as the outcome to be able to estimate ratio measures of association.

The function applied to the rate or risk in a regression analysis is referred to as the *link function* in the framework of generalized linear models underlying such analyses (see McCullagh and Nelder (1989) and Palta (2003) for theory and practical application). For example, linear regression would regress the risk or rate directly on exposure without any transformation, which is referred to as using the *identity link*. When the exposure is the only predictor in such a model, all link functions fit equally well and simply represent different ways to characterize the association. However, when several exposures or confounders are involved, or if the exposure is measured as a continuous or ordinal variable, some link functions and not others may require interaction or non-linear terms to improve the fit. The considerations in choosing the link function parallel those for choosing a measure of association as multiplicative or additive and as computed from rates, risks, or odds, discussed above (see Table 4.1).

Both traditional and regression estimation is briefly overviewed below, with more details provided in chapters ▸Regression Methods for Epidemiological Analysis and ▸Survival Analysis of this handbook.

4.4.6.1 Estimation and Adjustment for Confounding of Rate Ratios

Estimation of the rate or hazard ratio between exposed and non-exposed individuals can be based either on event counts (overall or in subgroups and/or subintervals of time) or on the time to event for each individual, where the time for subjects without events are entered as time to end of follow-up and are referred to as being *censored* (see chapter ▸Survival Analysis of this handbook).

In the first case, estimation can proceed directly by forming ratios of interest or by modeling the number of events on exposure by a generalized linear model. When ratios are formed directly as the ratio of the number of cases D_E divided by the person time at risk t_E, i.e., D_E/t_E, in those exposed and $D_{\bar{E}}/t_{\bar{E}}$ in those unexposed, the 95% confidence interval of the resulting rate ratio $HR = D_E/t_E/D_{\bar{E}}/t_{\bar{E}}$ is obtained as (Rothman and Greenland 1998, page 238)

$$[\exp(\ln(\widehat{HR}) - 1.96(1/D_E + 1/D_{\bar{E}})^{1/2}), \exp(\ln(\widehat{\mathrm{HR}}) + 1.96(1/D_E + 1/D_{\bar{E}})^{1/2})].$$

In either case, it is often necessary to adjust for confounding factors, including age and sex. When rate ratios are formed directly, the rates are generally adjusted by direct standardization (see Sect. 4.6.1) or by use of the *standardized mortality (*or *morbidity* or *incidence) ratio SMR* or *SIR* (see Sect. 4.6.1). The SMR and SIR have found wide application in investigations of the potential health effects of occupational exposures.

A common regression approach to estimating rate ratios requires information on event count and person time at risk for each subgroup, time interval and exposure level of interest. To obtain rate ratios from the regression requires that the logarithm of the mean number of events be modeled. This is referred to in the generalized linear model framework as using a *log link* function. The resulting regression equation is

$$\ln(h_i) = -\ln(t_i) + \beta_0 + \beta_E E_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \qquad (4.8)$$

where the subscript i indicates subject, $i = 1, \ldots. n$, and E_i is an indicator that equals 0 for the unexposed and 1 for the exposed. In this equation, β_0 is the logarithm of the rate per time unit for the unexposed with confounder values, $X_1, X_2, \ldots = 0$. Care should be taken to center confounders so that this intercept is meaningful. The quantity $\ln(t_i)$ is referred to as the offset and allows event counts over different size denominators to be used as the outcome variable. In the case when disease rates in a population are modeled, t_i are population sizes. The rate ratio for exposure adjusted for confounders $X_1, X_2, \ldots$ is obtained as exp (β_E). Differences in rate ratios across levels of X can easily be accommodated by the inclusion of interaction terms in the model. Inferences on the rate ratio follow from the standard error of the estimate $\hat{\beta}_E$ of β_E, which is approximately normally distributed with reasonable large sample sizes, so that a 95% confidence interval for the rate ratio is

$$\left[\exp\left(\hat{\beta}_E - 1.96\text{se}(\hat{\beta}_E)\right); \left(\exp(\hat{\beta}_E + 1.96\text{se}(\hat{\beta}_E)\right)\right].$$

The standard errors se$(\hat{\beta}_E)$ can be obtained from maximum-likelihood theory, assuming that the counts follow a *Poisson* or *negative binomial distribution.* The variance of the Poisson distribution equals the rate, while the negative binomial distribution allows for the variance being larger than the rate.

There are also several approaches available in most statistical software packages to adjust standard errors for so-called overdispersion. Overdispersion refers to variability in rates being larger than expected from a Poisson count process. For example, events may cluster in time, or there may be unmeasured characteristics of the population influencing the rate, so that the overall count arises from a mixture of different rates. An example of overdispersion (Palta 2003) arises in overall cancer rates because different cancers predominate for different ages and genders. One of the approaches to adjusting for overdispersion is to use a *robust* or *sandwich* estimator of the standard error of $\hat{\beta}_E$ available in software packages, such as PROC GENMOD in SAS (1999) that fit *generalized estimating equations* (Liang and Zeger 1986).

When the data consist of times to event for individuals, the rate ratio or hazard ratio can be estimated by techniques designed for *survival analysis* (e.g., see Hosmer and Lemeshow 1999 and chapter ▸Survival Analysis of this handbook).

Most parametrically specified survival distributions (i.e., distributions $S(t) = 1 - F(t)$, where F is the distribution of time to event) lead to hazard ratios $h_E(t)/h_{\bar{E}}(t)$ that vary over time. When the hazard ratio remains constant, this is referred to as proportional hazards. This property holds when the time to event follows the *exponential distribution*, so that the probability of avoiding an event up to time t is given by $S(t) = \exp(-ht)$ where h is a constant hazard rate and for the *Weibull distribution* $S(t) = \exp[(-ht)^{\gamma}]$ as long as γ is the same for the exposed and non-exposed groups. Models are sometimes fit that assume that the exponential distribution holds over short intervals, i.e., piecewise constant hazard rate. In these models, the hazard ratio is constant across short intervals but can be allowed to change over time. An exponential distribution for time to event leads to the Poisson distribution for number of events in a given time period.

In the situation of proportional hazards, estimation of the hazard ratio can proceed without specifying the actual survival distribution via the *Cox model*, where estimation is based on so-called partial likelihood (Cox 1972). The reason this works is that the actual level of the hazard rate cancels out; similarly to how the offset becomes part of the intercept in the regression model given by Eq. 4.8 above.

Table 4.3 Notation for a generic 2 × 2 table from cohort or case-control study

	Exposed	Unexposed
Diseased	a	b
Non-diseased	c	d

4.4.6.2 Estimation and Adjustment for Confounding for Risk Ratios

In a cohort study with a fixed follow-up time, the risk ratio can be estimated in a straightforward manner. From a 2×2 table (see Table 4.3) with cells *a, b, c, d,* where a is the number diseased and exposed, b is the number diseased and unexposed, c the number non-diseased and exposed, and d the number non-diseased and unexposed, the risk ratio is estimated by $\widehat{RR} = \{a/(a + c)\}/\{b/(b + d)\}$.

Statistical inference can be based on the approximate standard error (Katz et al. 1978) which can be estimated as $\mathrm{se}(\ln(\widehat{RR})) = \{a/(a+c)+d/b(b+d)\}^{1/2}$. In cases where follow-up time is not fixed, the risk ratio can be calculated from the individual risks estimated from the incidence rate or hazard rate. However, this is rarely done, as investigators tend to prefer the hazard ratio as the measure of association in such situations. The risk ratio can be estimated from case-control studies only when the ratio of sampling probabilities of cases and controls is known or by using the odds ratio (see Sect. 4.4.6.3) as an approximation.

Although standardization can be used either as direct standardization to adjust risks before forming ratios or as indirect standardization to compute the SMR (see Sect. 4.6.1) from risks in a reference population, it is often more appropriate to apply stratified analyses to adjust the risk ratio for confounders (see also Sect. 4.6.1). For example, a study of cancer risk in individuals exposed or not exposed to a risk factor may be stratified into age groups, or a study investigating outcomes in neonates may be stratified by birth weight. Stratum-specific risk ratio estimates can be calculated and then be combined for instance by the popular Mantel–Haenszel estimator that is known to have good properties. It is given by

$$\widehat{RR}_{\mathrm{MH}} = \sum (a_i(b_i + d_i)/n_i) \Big/ \sum (b_i(a_i + c_i)/n_i),$$

where the sums are across strata and n_i is the number of subjects in stratum i. This estimator is stable in small samples but has a larger standard error than the corresponding estimator from regression modeling. Formulas for the standard error are provided by Breslow and Day (1987) and Rothman and Greenland (1998).

From regression analysis, the risk ratio can be obtained as $\exp(\beta_E)$ from fitting the binary or *binomial* (grouped binary events) outcome to the model:

$$\ln(\pi_i) = \beta_0 + \beta_E E_i + \beta_1 X_{1i} + \ldots \tag{4.9}$$

This is a generalized linear model with error distribution reflecting each binary outcome being independent with variance $\pi_i(1 - \pi_i)$ and log link. Clearly, the log link is not ideal, as $\pi_i > 1$ can result from some exposure-confounder combinations. Nonetheless, this model tends to be reasonable with low risks. Maximum likelihood or generalized estimating equation fitting automatically provides large sample inference, with or without adjustment for deviations from the binomial error structure by robust standard errors. Deviations from binomial structure may result from clustering or correlation between events within subgroups or from multiple events per person (e.g., cavities in teeth when teeth are individually counted).

Another option for the link function when modeling the risk by a generalized linear model is the so-called complementary log-log link resulting in the model:

$$\ln(-\ln(1-\pi_i)) = \beta_0 + \beta_E E_i + \beta_1 X_{1i} + \ldots . \tag{4.10}$$

This model has the advantage of always estimating risks to be in the range 0–1. However, $\exp(\beta_E)$ is the rate ratio rather than the risk ratio.

4.4.6.3 Estimation and Adjustment for Confounding for Odds Ratios

In the traditional setting, the odds ratio in an unmatched case-control or cohort study is estimated from a 2×2 table (see Table 4.3) as $\widehat{OR} = ad/bc$. Inference can be based on exact methods, which historically were difficult to implement, but are now available in most statistical software packages, such as the SAS procedure PROC FREQ. With the exact approach, the confidence interval for the odds ratio is obtained from the *non-central hypergeometric distribution*. Over the years, many approximations to this interval have been developed, the most accurate of which is the *Cornfield approximation* (Cornfield 1956). Another, less accurate method is based on the approximate standard error of $\ln(\widehat{OR})$ known as the Woolf (1955) or *logit method*, where $\text{se}(\ln(\widehat{OR}))$ is calculated as $(1/a + 1/b + 1/c + 1/d)^{1/2}$. The logit method takes its name from being related to an approximation used for fitting logistic regression. Although the approximation has limited use for reporting final study results, it is useful to have an explicit approximation of the standard error for study planning purposes.

Stratified methods for estimating the odds ratio either build on taking a weighted average of the stratum specific log odds ratios, using the inverses of the logit method standard errors for each stratum as the weights, or using the Mantel-Haenszel stratified odds ratio estimator (Mantel and Haenszel 1959),

$$\widehat{OR}_{\text{MH}} = \left(\sum a_i c_i / n_i\right) \Big/ \left(\sum b_i d_i / n_i\right),$$

where the sums are across strata with tables as depicted in Table 4.3 for each stratum and n_i is the number of subjects in stratum i. This odds ratio estimator has been shown to have excellent properties even when strata are very small (Birch 1964; Breslow 1981; Breslow and Day 1980, Chaps. 4–5; Landis et al. 1978, 2000; Greenland 1987; Robins and Greenland 1989). The confidence interval for a stratified odds ratio can be obtained by exact methods or by the approximation of Miettinen (1976) where $\text{se}(\ln(\widehat{OR}))$ is calculated as $\ln(\widehat{OR}_{\text{MH}})/\chi_{\text{MH}}$. Here χ_{MH} is the square root of the *Mantel–Haenszel stratified chi-square test* used to test the null hypothesis that the odds ratio equals one (Mantel and Haenszel 1959). This test statistic is computed as

$$\chi^2_{\mathrm{MH}} = \Sigma[a_i - (a_i + b_i)(a_i + c_i)/n_i]^2/[(a_i + b_i)(a_i + c_i)(d_i + b_i)(d_i + c_i)/(n_i^2(n_i - 1))].$$

The 95% confidence interval for the odds ratio is then given by

$$[\exp(\ln(\widehat{OR}) - 1.96(\ln(\widehat{OR}_{\mathrm{MH}})/\chi_{\mathrm{MH}}), \exp(\ln(\widehat{OR}) + 1.96(\ln(\widehat{OR}_{\mathrm{MH}})/\chi_{\mathrm{MH}}))].$$

A special case of stratification occurs when data are pair matched, such that each case is matched to a control, e.g., based on kinship or neighborhood. In this case, the Mantel–Haenszel odds ratio estimator becomes m_{++}/m_{--} , where m_{++} is the number of pairs (matched sets) where both the case and the control are exposed, and m_{--} is the number of pairs where neither is exposed. Breslow and Day (1980, Chap. 7) provide additional formulas for the situation when several controls are matched to each case. Confidence intervals can again be obtained by exact formulas (Breslow and Day 1980, Chap. 7). It is well known that although matched studies are not technically confounded by the factors matched on because cases and controls are balanced on these, odds ratios based on the matched formula are larger than odds ratios not taking the matching into account. This is an example of the point made above regarding the non-collapsibility of odds ratios in Sect. 4.4.3.4. We discuss this phenomenon further in the logistic model framework below.

Increasingly, logistic regression is used for the estimation of odds ratios from clinical trials, cohort and case-control studies. Logistic regression fits the equation

$$\ln(\pi_i/(1 - \pi_i)) = \beta_0 + \beta_E E_i + \beta_1 X_{1i} + \ldots \tag{4.11}$$

with E_i denoting the exposure status and X_{1i}, $X_{2i}, \ldots$ the confounder variables of individual i. For a cohort study, β_0 is $\ln(\pi_i/(1 - \pi_\iota))$ for an unexposed individual with all confounders equal to 0. For such a person, then, the risk of disease $\pi_i = \exp(\beta_0)/[1 + \exp(\beta_0)]$. In a case-control study, the intercept in Eq. 4.11 is $\beta_0 = \beta_{0,\mathrm{cohort}} + \ln(P_1/P_0)$, with P_1 and P_0 the probabilities for being sampled into the study for cases and controls, respectively. We see again that risk can be estimated from a case-control study only when the sampling scheme of cases and controls is known. The odds ratio for exposure, adjusted for confounders, is $\exp(\beta_E)$.

In the generalized linear model framework, Eq. 4.11 is said to use the *logit link*, where the logit function is defined as $g(\pi) = \ln(\pi/(1 - \pi))$. The logit link is the one that follows most naturally from the mathematical formulation of the binomial distribution (McCullagh and Nelder 1989) and is referred to as the *canonical link*, whereas the log is the canonical link for rates. Just as for other generalized linear models, maximum-likelihood-based and robust standard errors are available, with the latter taking into account clustering of events. It should be noted, however, that generalizations of logistic regression to the longitudinal or clustered setting by generalized estimating equations (GEE) do not work for case-control studies

(Neuhaus and Jewell 1990; for a general introduction to GEE see e.g. chapter ►Generalized Estimating Equations of this handbook).

One point that has received considerable attention for clustered or matched data is the difference in estimated odds ratios between so-called marginal logistic models usually fit via GEE and so-called cluster specific models that explicitly model random intercepts for each cluster or matched set (Neuhaus et al. 1991). This difference is another manifestation of the non-collapsibility of odds ratios, mentioned in Sect. 4.4.3.4, and arises even when there is no confounding by cluster-specific factors (see, e.g., Hanley et al. 2003). Some further insight into this problem can be generated by realizing that logistic regression arises from dichotomizing a continuous underlying latent variable with a logistic distribution (conditional on covariates). The logistic regression coefficients depend on the standard deviation of that continuous latent variable, and the difference between marginal and cluster specific coefficients is a reflection of inherently choosing the dependence to be on overall versus within cluster standard deviations, respectively (Palta et al. 1997; Palta and Lin 1999). General problems of using standard deviation units of outcomes have been pointed out (Greenland et al. 1986) and are inherent and unavoidable in the use of logistic regression for clustered data.

Matched data can also be analyzed by *conditional logistic regression* that fits the model

$$\ln(\pi_{ji}/(1-\pi_{ji})) = \beta_{0j} + \beta_E E_{ji} + \beta_1 X_{1ji} + \ldots \tag{4.12}$$

for individual i in the matched set j. Estimation of β_E and β_1, ... is based on algorithms that compare individuals only within and not between matched sets. For example, for matched pairs, estimation is based on differences in exposure and confounders. Equation 4.12 is a cluster-specific model but differs from random effects models in the fact that the intercepts β_{0j} can reflect true confounders, while intercepts in the random effects models are assumed to be independent of risk factors. The algorithms for conditional logistic regression do not actually estimate the matched set-specific intercepts β_{0j} that cancel out. All variables that do not vary within matched sets are automatically absorbed into β_{0j} although interactions of such variables with those that vary within set can be included in the model. While the conditional logistic regression model is usually fit by large sample methods, such as maximum likelihood, so-called exact procedures have also become available (e.g., Mehta et al. 2000). Again, taking matching into account in the analysis results in larger coefficients than those of the unmatched model (4.11). When all matching variables are explicit (such as age and sex), they can be directly entered as covariates in Eq. 4.11.

4.4.6.4 Estimation and Adjustment for Confounding for Rate Differences

Regression estimation of the rate difference with and without adjustment for confounders can be done in the generalized linear model framework by specifying the identity link function, resulting in linear regression of the rates with variance arising from the Poisson distribution. Overdispersion can be handled the same way as for ratios. However, unequal time intervals cannot be as easily accommodated with the identity link. Instead, weighted ordinary regression of observed rates can be employed, where inverse variance weights automatically account for the interval length (Breslow and Day 1987, Chap. 4).

4.5 Measures of Impact

Measures of impact are used to assess the contribution of one or several exposures to the occurrence of incident cases at the population level. Thus, they are useful in public health to weigh the impact of exposure on the burden of disease occurrence and assess potential prevention programs aimed at reducing or eliminating exposure in the population. They are sometimes referred to as measures of *potential* impact to convey the notion that the true impact at the population level may be different from that reflected by these measures except under very specific conditions (see Sect. 4.5.1.4). The most commonly used measure of impact is the attributable risk. This measure is presented in some detail below. Then, other measures are briefly described. Table 4.4 provides an overview of measures of impact discussed in this chapter.

4.5.1 Attributable Risk

4.5.1.1 Definition

The term "attributable risk" (*AR*) was initially introduced by Levin in 1953 as a measure to quantify the impact of smoking on lung cancer occurrence. Gradually, it has become a widely used measure to assess the consequences of an association between an exposure and a disease at the population level. It is defined as the following ratio:

$$AR = \{\Pr(D) - \Pr(D|\overline{E})\}/\Pr(D). \tag{4.13}$$

The numerator contrasts the probability of disease, $\Pr(D)$, in the population, which may have some exposed, E, and some unexposed, $\overline{E}$, individuals, with the hypothetical probability of disease in the same population but with all exposure eliminated $\Pr(D|\overline{E})$. Thus, it measures the additional probability of disease in the population that is associated with the presence of an exposure in the population, and *AR* measures the corresponding proportion. Probabilities in Eq. 4.13 will usually refer to disease risk although, depending on the context, they may be replaced with incidence rates.

Unlike measures of association (see Sect. 4.4), *AR* depends both on the strength of the association between exposure and disease and the prevalence of exposure in the population, p_E. This can be seen, for instance, through rewriting *AR* from Eq. 4.13. Upon expressing $\Pr(D)$ as

$$\Pr(D|E)p_E + \Pr(D|\overline{E})p_{\bar{E}} \text{ with } p_{\bar{E}} = 1 - p_E,$$

both in the numerator and the denominator, and noting that

$$\Pr(D|E) = \text{RR} \times \Pr(D|\overline{E}),$$

the term $\Pr(D|\overline{E})$ cancels out and AR is obtained as (Cole and MacMahon 1971; Miettinen 1974)

$$AR = \{p_E(RR - 1)\}/\{1 + p_E(RR - 1)\}, \tag{4.14}$$

a function of both the prevalence of exposure in the population, p_E, and the rate ratio or relative risk, RR.

An alternative formulation underscores this joint dependency in yet another manner. Again, upon expressing $\Pr(D)$ as

$$\Pr(D|E)p_E + \Pr(D|\overline{E})p_{\bar{E}} \text{ with } p_{\bar{E}} = 1 - p_E$$

and noting that

$$\Pr(D|E) = \text{RR} \times \Pr(D|\overline{E}),$$

the numerator in Eq. 4.13 can be rewritten as

$$p_E \Pr(D|E) - p_E \Pr(D|E)/RR.$$

From using Bayes' theorem to express $\Pr(D|E)$ as $\Pr(E|D)\Pr(D)/p_E$, it then becomes equal to

$$\Pr(D)p_{E|D}(1 - 1/RR),$$

after simple algebra. This yields (Miettinen 1974)

$$AR = p_{E|D}(RR - 1)/RR, \tag{4.15}$$

a function of the prevalence of exposure in diseased individuals, $p_{E|D}$, and the rate ratio or relative risk, RR.

A high relative risk can correspond to a low or high AR depending on the prevalence of exposure, which leads to widely different public health consequences. One implication is that, portability is not a usual property of AR, as the prevalence of exposure may vary widely among populations that are separated in time or location. This is in contrast with measures of association such as the relative risk or rate ratio which are more portable from one population to another, as the strength of the association between disease and exposure might vary little among populations (unless strong interactions with environmental or genetic factors are present). However, portability of RR can be questioned as well in the case of imperfect specificity of exposure assessment, since misclassification of non-exposed subjects as exposed will bias RR toward unity, which will affect differentially RR estimates in various populations depending on their exposure prevalence. This is not a problem with AR, which is not affected by imperfect specificity of exposure assessment.

Table 4.4 Measures of impact discussed in this chapter

Measures	Range	Definition[a]	Usual interpretation (s)[b]
Attributable risk (*AR*)	$-\infty$ to 1 0 to 1 for risk factor	1. $\{\Pr(D) - \Pr(D\|\overline{E})\}/\Pr(D)$ 2. $\{p_E(RR-1)\}/\{1 + p_E(RR-1)\}$ 3. $p_{E\|D}(RR-1)/RR$	Proportion of disease cases in the population attributable to exposure Proportion of disease cases in the population potentially preventable by eliminating exposure
Attributable risk among the exposed (AR_E)	$-\infty$ to 1 0 to 1 for risk factor	1. $\{\Pr(D\|E) - \Pr(D\|\overline{E})\}/\Pr(D\|E)$ 2. $(RR - 1)/RR$	Proportion of disease cases among the exposed attributable to exposure
Sequential attributable risk	0 to 1 for risk factor	Contributions of a given exposure to the joint attributable risk to several exposures for a given order of exposures	Proportion of disease cases in the population potentially preventable by eliminating a given exposure when several exposures are removed in a given sequence
Partial attributable risk	0 to 1 for risk factor	Average contribution of a given exposure to the joint attributable risk to several exposures over all possible exposure orders	Average proportion of disease cases in the population potentially preventable by eliminating a given exposure when several exposures are removed in sequence over all possible orders of removal
Prevented (preventable) fraction (PF)	$-\infty$ to 1 0 to 1 for protective factor	1. $\{\Pr(D\|\overline{E}) - \Pr(D)\}/\Pr(D\|\overline{E})$ 2. $p_E(1 - RR)$	Proportion of disease cases averted ("prevented fraction") in relation to the presence of a protective exposure or intervention in the population Proportion of cases that could be potentially averted ("preventable fraction") if a protective exposure or intervention were introduced de novo in the population

Generalized impact fraction	$-\infty$ to 1 0 to 1 for risk factor and modified distribution with lowering of exposure	$\{\Pr(D) - \Pr^*(D)\}/\Pr(D)$	Proportion of disease cases potentially averted (fractional reduction of disease occurrence) from changing the current distribution of exposure in the population to some modified distribution
Person-years of life lost (PYLL)	$\geq$0 for risk factor (person-years)	Difference between current life expectancy and life expectancy with exposure removed at the population level	Person-time of life lost at the population level attributable to exposure
Average potential years of life lost (PYLL)	$\geq$0 for risk factor (years)	Average difference per exposed person between current life expectancy and life expectancy with exposure removed	Average loss of life expectancy per person attributable to exposure

[a] $\Pr(D)$, $\Pr(D|\overline{E})$, $\Pr(D|E)$, and $\Pr^*(D)$ denote probabilities of disease (disease risks), namely, the overall probability of disease in the population, the probability of disease in the population with all exposure eliminated, the probability of disease among exposed individuals, and the overall probability of disease under a modified distribution of exposure, respectively. Alternatively, they may refer to disease rates depending on the context. The terms p_E and $p_{E|D}$ respectively refer to the overall exposure prevalence in the population and the exposure prevalence in the diseased individuals. The term *RR* refers to risk or rate ratios for exposed relative to unexposed individuals.

[b] Interpretations subject to conditions (see text, e.g., Sect. 4.5.1.4)

4.5.1.2 Range

When the exposure considered is a risk factor ($RR > 1$), it follows from the above definition that AR lies between 0 and 1. Therefore, it is very often expressed as a percentage. AR increases both with the strength of the association between exposure and disease measured by RR and with the prevalence of exposure in the population. A prevalence of 1 (or 100%) yields a value of AR equal to the attributable risk among the exposed, i.e., $(RR - 1)/RR$. AR approaches 1 for an infinitely high RR provided the exposure is present in the population (i.e., non-null prevalence of exposure).

AR takes a null value when either there is no association between exposure and disease ($RR = 1$) or there are no exposed subjects. Negative AR values are obtained for a protective exposure ($RR < 1$). In this case, AR varies between 0 and $-\infty$, a scale on which AR lacks a meaningful interpretation. One solution is to reverse the coding of exposure (i.e., interchange exposed and unexposed categories) to go back to the situation of a positive AR, sometimes called the preventable fraction in this case (Benichou 2000c; Greenland 1987; Last 1983). Alternatively, one can consider a different parameter, namely, the prevented fraction (see Sect. 4.5.4).

4.5.1.3 Synonyms

Some confusion in the terminology arises from the reported use of as many as 16 different terms in the literature to denote attributable risk (Gefeller 1990, 1995). However, a literature search by Uter and Pfahlberg (1999) found some consistency in terminology usage, with "attributable risk" and "population attributable risk" (MacMahon and Pugh 1970) the most commonly used terms by far, followed by "etiological fraction" (Miettinen 1974). Other popular terms include "attributable risk percentage" (Cole and MacMahon 1971), "fraction of etiology" (Miettinen 1974), and "attributable fraction" (Greenland and Robins 1988; Last 1983; Ouellet et al. 1979; Rothman and Greenland 1998, Chap. 4).

Moreover, additional confusion may originate in the use by some authors (MacMahon and Pugh 1970; Markush 1977; Schlesselman 1982) of the term "attributable risk" to denote a measure of association, the excess incidence, that is the difference between the incidence rates in exposed and unexposed subjects (see Sect. 4.4.4.2). Context will usually help the readers detect this less common use.

4.5.1.4 Interpretation and Usefulness

While measures of association such as the rate ratio and relative risk are used to establish an association in etiological research, AR has a public health interpretation as a measure of the disease burden attributable or at least related to one or several exposures. Consequently, AR is used to assess the potential impact of prevention programs aimed at eliminating exposure from the population. It is often thought of as the fraction of disease that could be eliminated if exposure could be totally removed from the population.

However, this interpretation can be misleading because, for it to be strictly correct, the three following conditions have to be met (Walter 1976). First, estimation of AR has to be unbiased (see Sect. 4.5.1.7). Second, exposure has to be causal rather than merely associated with the disease. Third, elimination of exposure has to be without any effect on the distribution of other risk factors. Indeed, as it might be difficult to alter the level of exposure to one factor independently of other risk factors, the resulting change in disease load might be different from the *AR* estimate. For these reasons, various authors elect to use weaker definitions of *AR*, such as the proportion of disease that can be related or linked, rather than attributable, to exposure (Miettinen 1974).

A fundamental problem regarding causality has been discussed by Greenland and Robins (1988) and Robins and Greenland (1989) who considered the proportion of disease cases for which exposure played an etiological role, i.e., cases for which exposure was a component cause of disease occurrence. They termed this quantity the etiological fraction and argued that it was a more relevant measure of impact than *AR*. Rothman and Greenland (1998, Chap. 4) argued that *AR* and the etiological fractions are different quantities using logical reasoning regarding causality and the fact that disease occurrence may require several component causal factors rather than one. The main problem with the etiological fraction is that it is usually impossible to distinguish exposed cases for whom exposure played an etiological role from those where exposure was irrelevant. As a consequence, estimating the etiological fraction will typically require non-identifiable biological assumptions about exposure actions and interactions to be estimable (Cox 1984, 1985; Robins and Greenland 1989; Seiler 1987). Thus, despite its limitations, *AR* remains a useful measure to assess the potential impact of exposure at the population level and can serve as a suitable guide in practice to assess and compare various prevention strategies.

Several authors have considered an interpretation of *AR* in terms of etiological research. The argument is that if an *AR* estimate is available for several risk factors jointly, then its complement to 1, $1 - AR$, must represent a gauge of the proportion of disease cases not explained by the risk factors used in estimating *AR*. Hence, $1 - AR$ would represent the proportion of cases attributable to other (possibly unknown) risk factors. For instance, it was estimated that the *AR* of breast cancer was 41% for late age at first birth, nulliparity, family history of breast cancer, and higher socioeconomic status, which suggested that at least 59% of cases had to be attributable to other risk factors (Madigan et al. 1995). A similar type of reasoning was used in several well-known reports of estimated percentages of cancer death or incidence attributable to various established cancer risk factors (e.g., smoking, diet, occupational exposure to carcinogens…). Some of these reports conveyed the impression that little remained unexplained by factors other than the main established preventable risk factors and that cancer was a mostly preventable illness (Colditz et al. 1996, 1997; Doll and Peto 1981; Henderson et al. 1991; Ames et al. 1995). Such interpretation has to be taken with great caution since *AR*s for different risk factors may add to more than 100% because multiple exposures are usually possible (e.g., smoking and occupational exposure to asbestos). Moreover, this interpretation can be refuted on the basis of logical arguments regarding the fact that disease occurrence may require more than one causal factor (see Rothman and Greenland 1998, Chap. 2). Furthermore, one can note that once a new risk factor is considered, the joint unexposed reference category changes from lack of exposure to all previously considered risk factors to lack of exposure to those risk factors *and* the new risk factor (Begg 2001). Because of this change in the reference category, the *AR* for the new risk factor may surpass the quantity $1-AR$ for previously considered risk factors. Thus, while it is useful to know that only 41% of breast cancer cases can be attributed to four established risk factors in the above example, it is entirely conceivable that new risk factors of breast cancer may be elicited which yield an *AR* of more than 59% by themselves in the above example.

4.5.1.5 Properties

AR has two main properties. First, *AR* values greatly depend on the definition of the reference level for exposure (unexposed or baseline level). A more stringent definition of the reference level corresponds to a larger proportion of subjects exposed, and, as one keeps depleting the reference category from subjects with higher levels of risk, *AR* values and estimates keep rising. This property has a major impact on *AR* estimates as was illustrated by Benichou (1991) and Wacholder et al. (1994). For instance, Benichou (1991) found that the *AR* estimate of esophageal cancer for an alcohol consumption greater or equal to 80 g/day (reference level of $0-79$ g/day) was 38% in the Ille-et-Vilaine district of France and increased dramatically to 70% for an alcohol consumption greater or equal to 40 g/day (i.e., using the more restrictive reference level $0-39$ g/day) (see Example 4.3 below). This property plays a role whenever studying a continuous exposure with a continuous gradient of risk and when there is no obvious choice of threshold. Therefore, *AR* estimates must be reported with reference to a clearly defined baseline level in order to be validly interpreted.

Example 4.3.

A case-control study of esophageal cancer conducted in the Ille-et-Vilaine district of France included 200 cases and 775 controls selected by simple random sampling from electoral lists (Tuyns et al. 1977). The assessment of associations between alcohol consumption and smoking with esophageal cancer has been the focus of detailed illustration by Breslow and Day (1980) who presented various approaches to odds ratio estimation with or without adjustment for age. As in previous work (Benichou 1991), four levels of alcohol consumption (0–39, 40–79, 80–119, and 120+ g/day) are considered here as well as three levels of smoking (0–9, 10–29, 30+ g/day) and three age groups (25–44, 45–54, 55+ years). There were 29, 75, 51, and 45 cases with respective alcohol consumptions of 0–39, 40–79, 80–119, and 120+ g/day. Corresponding numbers of controls were 386, 280, 87, and 22, respectively. The first reference level considered, 0–79 g/day, included 104 cases and 666 controls, leaving 96 cases and 109 controls in the exposed (i.e., 80+ g/day) category (see Table 4.5). The corresponding crude (unadjusted) odds ratio was estimated as $(96 \times 666)/(104 \times 109) = 5.6$ (see Sect. 4.4.6.3). Using methods described below (see Sects. 4.5.1.6 and 4.5.1.7), the crude *AR* estimate was 39.5% for alcohol consumption, and the age- and smoking-adjusted *AR* estimates were close to 38%. The second reference level considered, $0-39$ g/day, was more restrictive and included only 29 cases and 286 controls, leaving 171 cases and 489 controls in the exposed (i.e., 40+ g/day) category (see Table 4.5). The corresponding crude odds ratio was estimated as $(171 \times 386)/(29 \times 389) = 5.9$ (see Sect. 4.4.6.3). Using methods described below (see Sects. 4.5.1.6 and 4.5.1.7), the crude *AR* estimate was 70.9%, and adjusted *AR* estimates were in the range 70–72%. The marked increase mainly resulted from the much higher proportion of subjects exposed with the more restrictive definition of the reference category (63% instead of 14% of exposed controls).

Table 4.5 Numbers of cases and controls in the reference and exposed categories of daily alcohol consumption according to two definitions of the reference category – data from a case-control study of esophageal cancer (From Tuyns et al. 1977)

More restrictive definition of reference category (0–39 g/day)				Less restrictive definition of reference category (0–79 g/day)			
	Reference category (0–39 g/day)	Exposed category (40+ g/day)	Total		Reference category (0–79 g/day)	Exposed category (80+ g/day)	Total
Cases	29	171	200	Cases	104	96	200
Controls	386	389	775	Controls	666	109	775
Total	315	660	975	Total	770	205	975

The second main property is distributivity. If several exposed categories are considered instead of just one, then the sum of the category-specific *AR*s equals the overall *AR* calculated from combining those exposed categories into a single one, regardless of the number and the divisions of the original categories (Benichou 1991; Wacholder et al. 1994; Walter 1976), provided the reference category remains the same. This property applies strictly to crude *AR* estimates and to adjusted *AR* estimates calculated on the basis of a saturated model including all possible interactions (Benichou 1991). It applies approximately to adjusted estimates not based on a saturated model (Wacholder et al. 1994). Thus, if an overall *AR* estimate is the focus of interest, there is no need to break the exposed category into several mutually exclusive categories, even in the presence of a gradient of risk with increasing level of exposure. Of course, if the impact of a partial removal of exposure is the question of interest, retaining detailed information on the exposed categories will be necessary (Greenland 2001).

Example 4.3. (Continued)
For the more restrictive definition of the reference category of daily alcohol consumption (0–39 g/day), the crude *AR* was estimated at 70.9%. The separate contributions of categories 40–79, 80–119, and 120+ g/day were 27.0%, 22.2%, and 21.7%, summing to the same value 70.9%. Similarly, for the less restrictive definition of the reference category (0–79 g/day), the crude *AR* was estimated at 39.5% and the separate contributions of categories 80–119 g/day and 120+ g/day were 18.7% and 20.8%, summing to the same value 39.5%.

4.5.1.6 Estimability and Basic Principles of Estimation

AR can be estimated from cohort studies since all quantities in Eqs. 4.13, 4.14, and 4.15 are directly estimable from cohort studies. *AR* estimates can differ depending on whether rate ratios, risk ratios, or odds ratios are used but will be numerically close for rare diseases. For case-control studies, exposure-specific incidence rates or risks are not available unless data are complemented with follow-up or population-based data (see Sect. 4.2.2). Thus, one has to rely on odds ratio estimates, use Eq. 4.14, and estimate p_E from the proportion exposed in the controls, making the rare-disease assumption also involved in estimating odds ratios rather than relative risks. For crude *AR* estimation, the estimate of the odds ratio is taken as ad/bc and that of p_E as $c/(c+d)$, where, as in Sect. 4.4, a, b, c, and d, respectively, denote the numbers of exposed cases, unexposed cases, exposed controls, and unexposed controls. Alternatively, one can use Eq. 4.15, in which the quantity $p_{E|D}$ can be directly estimated from the diseased individuals (cases) as $a/(a+b)$ and *RR* can be estimated from the odds ratio again as ad/bc. Using either equation, the resulting point estimate is given by $(ad - bc)/\{d(a+b)\}$.

Variance estimates of crude *AR* estimates are based on applying the delta method (Rao 1965). For instance, an estimate of the variance for case-control data is given by the quantity

$$\mathrm{var}(\widehat{AR}) = b(c+d)\{ad(c+d) + bc(a+b)\}/\{d^3(a+b)^3\}.$$

Various $(1-\alpha)\%$ confidence intervals for *AR* have been proposed that can be applied to all epidemiological designs once point, and variance estimates are obtained. They include standard confidence intervals for *AR* based on the untransformed *AR* point estimate, namely

$$\widehat{AR} \pm z_{1-\alpha/2}\mathrm{se}(\widehat{AR});$$

AR confidence intervals based on the log-transformed variable $\ln(1-AR)$, namely,

$$1-(1-\widehat{AR})\left\lfloor \exp\left\{\pm z_{1-\alpha/2}\mathrm{se}(\widehat{AR})/(1-\widehat{AR})\right\}\right\rfloor \text{ (Walter 1975);}$$

as well as confidence intervals based on the logit-transformed variable $\ln\{AR/(1-AR)\}$, namely,

$$\left\{1+\left\{(1-\widehat{AR})/\widehat{AR}\right\}\left(\exp\left[\pm z_{1-\alpha/2}\mathrm{se}(\widehat{AR})\Big/\left\{\widehat{AR}(1-\widehat{AR})\right\}\right]\right)\right\}^{-1}$$

(Leung and Kupper 1981).

In the previous formulae, $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$th percentile of the standard normal distribution, $\widehat{AR}$ denotes the *AR* point estimate, and $\mathrm{se}(\widehat{AR})$ denotes its corresponding standard error estimate. Whittemore (1982) noted that the log transformation yields a wider interval than the standard interval for $AR > 0$. Leung and Kupper (1981) showed that the interval based on the logit transform is narrower than the standard interval for values of *AR* strictly between 0.21 and 0.79, whereas the reverse holds outside this range for positive values of *AR*. While the coverage probabilities of these intervals have been studied in some specific situations and partial comparisons have been made, no general studies have been performed to determine their relative merits in terms of coverage probability.

Detailed reviews of estimability and basic estimation of *AR* for various epidemiological designs can be found in Walter (1976) and Benichou (2000b, 2001) who provide explicit formulae for $\widehat{AR}$ and $\mathrm{se}(\widehat{AR})$ for cohort and case-control designs.

Example 4.3. (Continued)
For the more restrictive definition of the reference category of daily alcohol consumption (0–79 g/day), the crude *AR* estimate was obtained as

$$(171 \times 386 - 29 \times 389)/(386 \times 200) = 0.709$$

or 70.9%. Its variance was estimated as

$$29 \times 775 \times (171 \times 386 \times 775 + 29 \times 389 \times 200)/(386^3 \times 200^3) = 0.00261,$$

yielding a standard error estimate of 0.051 or 5.1%. The corresponding 95% confidence intervals for AR are given by 60.9–80.9% (no transformation), 58.9–79.4% (log transformation), and 60.0–79.8% (logit transformation), very similar to each other in this example.

4.5.1.7 Adjusted Estimation

As is the case for measures of association, unadjusted (or crude or marginal) AR estimates may be inconsistent (Miettinen 1974; Walter 1976, 1980, 1983). The precise conditions under which adjusted AR estimates that take into account the distribution and effect of other factors will differ from unadjusted AR estimates that fail to do so were worked out by Walter (1980). If E and X are two dichotomous factors taking levels 0 and 1, and if one is interested in estimating the AR for exposure E, then the following applies. The adjusted and unadjusted AR estimates coincide (i.e., the crude AR estimate is unbiased) if and only if (a) E and X are such that $\Pr(E = 0, X = 0) \times \Pr(E = 1, X = 1) = \Pr(E = 0, X = 1) \times \Pr(E = 1, X = 0)$, which amounts to the independence of their distributions or (b) exposure to X alone does not increase disease risk, namely, $\Pr(D|E = 0, X = 1) = \Pr(D|E = 0, X = 0)$. When considering one (or several) polychotomous factor(s) X forming J levels ($J > 2$), conditions (a) and (b) can be extended to a set of analogous sufficient conditions. Condition (a) translates into a set of $J(J-1)/2$ conditions for all pairs of levels j and j' of X, amounting to an independent distribution of E and all factors in X. Condition (b) translates into a set of $J - 1$ conditions stating that in the absence of exposure to E, exposure to any of the other factors in X, alone or in combination, does not increase disease risk.

The extent of bias varies according to the severity of the departure from conditions (a) and (b) above. Although no systematic numerical study of the bias of unadjusted AR estimates has been performed, Walter (1980) provided a revealing example of a case-control study assessing the association between alcohol, smoking, and oral cancer. In that study, severe positive bias was observed for crude AR estimates, with a very large difference between crude and adjusted AR estimates both for smoking (51.3% vs. 30.6%, a 20.7 difference in percentage points and 68% relative difference in AR estimates) and alcohol (52.2% vs. 37.0%, a 15.2% absolute difference and 48% relative difference). Thus, the prudent approach must be to adjust for factors that are suspected or known to act as confounders in a similar fashion as for estimating measures of associations.

Two simple adjusted estimation approaches discussed in the literature are inconsistent. The first approach was presented by Walter (1976) and is based on a factorization of the crude risk ratio into two components similar to those in Miettinen's earlier derivation (Miettinen 1972). In this approach, a crude AR estimate is obtained under the assumption of no association between exposure and disease (i.e., values of RR or the odds ratio are taken equal to 1 separately for each level of confounding). This term reflects the *AR* only due to confounding factors since it is obtained under the assumption that disease and exposure are not associated. By subtracting this term from the crude *AR* estimate that ignores confounding factors and thus reflects the impact of both exposure and confounding factors, what remains is an estimate of the *AR* for exposure adjusted for confounding (Walter 1976). The second approach is based on using Eq. 4.14 and plugging in a common adjusted *RR* estimate (odds ratio estimate in case-control studies), along with an estimate of p_E (Cole and MacMahon 1971; Morgenstern and Bursic 1982). Both approaches, while intuitively appealing, were shown to be inconsistent (Ejigou 1979; Greenland and Morgenstern 1983; Morgenstern and Bursic 1982), and, accordingly, very severe bias was exhibited in simulations of cross-sectional and cohort designs (Gefeller 1995).

By contrast, two adjusted approaches based on stratification yield valid estimates. The Mantel–Haenszel approach consists in plugging-in an estimate of the common adjusted *RR* (odds ratio in case-control studies) and an estimate of the prevalence of exposure in diseased individuals, $p_{E|D}$, in Eq. 4.15 in order to obtain an adjusted estimate of *AR* (Greenland 1984, 1987; Kuritz and Landis 1987, 1988a, b). In doing so, it is possible to adjust for one or more polychotomous factors forming J levels or strata. While several choices are available for a common adjusted *RR* or odds ratio estimator, a usual choice is to use a Mantel–Haenszel estimator of *RR* in cohort studies (Kleinbaum et al. 1982, Chaps. 9 and 17; Landis et al. 2000; Rothman and Greenland 1998, Chaps. 15–16; Tarone 1981) or odds ratio in case-control studies (Breslow and Day 1980, Chaps. 4–5; Kleinbaum et al. 1982, Chaps. 9, 17; Landis et al. 2000; Mantel and Haenszel 1959; Rothman and Greenland 1998, Chaps. 15–16) (see Sect. 4.4.6). For this reason, the term “Mantel–Haenszel approach” has been proposed to denote this approach to adjusted *AR* estimation (Benichou 1991). When there is no interaction between exposure and factors adjusted for, Mantel–Haenszel type estimators of *RR* or odds ratio have favorable properties, as they combine lack of (or very small) bias even for sparse data (e.g., individually matched case-control data) and good efficiency except in extreme circumstances (Birch 1964; Breslow 1981; Breslow and Day 1980, Chaps. 4–5; Landis et al. 1978, 2000). Moreover, variance estimators are consistent even for sparse data (“dually consistent” variance estimators) (Greenland 1987; Robins and Greenland 1989). Simulation studies of cohort and case-control designs (Gefeller 1992; Greenland 1987; Kuritz and Landis 1988a, b) showed that adjusted *AR* estimates are little affected by small-sample bias when there is no interaction between exposure and adjustment factors but can be misleading if such interaction is present.

Example 4.3. (Continued)
In order to control for age and smoking, nine strata (joint categories) of smoking × age have to be considered. The Mantel–Haenszel odds ratio estimate can be calculated from quantities a_j, b_j, c_j, and d_j that respectively denote the numbers of exposed cases, unexposed cases, exposed controls, and unexposed controls in stratum j, using the methods in Sect. 4.4.6.3. With the more restrictive definition of the reference category for daily alcohol consumption, the Mantel–Haenszel odds ratio was estimated at 6.2, thus

slightly higher than the crude odds ratio of 5.9. Combined with an observed proportion of exposed cases of $171/200 = 0.855$, this resulted in an adjusted *AR* estimate of $0.855\times(6.2-1)/6.2 = 0.716$ or 71.6% using Eq. 4.15 (allowing for rounding error), slightly higher than the crude *AR* estimate of 70.9%. The corresponding estimate of the standard error was 5.1%.

The weighted-sum approach also allows adjustment for one or more polychotomous factors forming J levels or strata. The *AR* is written as a weighted sum over all strata of stratum-specific *AR*s, i.e., $\sum_{j=1}^{J} w_j AR_j$ (Walter 1976; Whittemore 1982, 1983). Using crude estimates of AR_j separately within each stratum j and setting weights w_j as proportions of diseased individuals (cases) yields an asymptotically unbiased estimator of *AR*, which can be seen to be a maximum-likelihood estimator (Whittemore 1982). This choice of weights defines the "case-load method." The weighted-sum approach does not require the assumption of a common relative risk or odds ratio. Instead, the relative risks or odds ratios are estimated separately for each adjustment level with no restrictions placed on them, corresponding to a fully saturated model for exposure and adjustment factors (i.e., a model with all interaction terms present). From these separate relative risk or odds ratio estimates, separate *AR* estimates are obtained for each level of adjustment. Thus, the weighted-sum approach not only accounts for confounding but also for interaction. Simulation studies of cohort and case-control designs (Gefeller 1992; Kuritz and Landis 1988a, b; Whittemore 1982) show that the weighted-sum approach can be affected by small sample bias, sometimes severely. It should be avoided when analyzing sparse data and should not be used altogether for analyzing individually matched case-control data.

Example 4.3. (Continued)
As with the Mantel–Haenszel approach, nine strata (joint categories) of smoking × age have to be considered in order to control for age and smoking. In each stratum separately, an *AR* estimate is calculated using the methods for crude *AR* estimation (see Sect. 4.5.1.6). For instance, among heavy smokers (30+ g/day) in age group 65+ years, there were 15 exposed cases, 5 unexposed cases, 4 exposed controls, and 6 unexposed controls, yielding an odds ratio estimate of 4.5 and an *AR* estimate of 58.3%. The corresponding weight was $20/200 = 0.1$, so that the contribution of this stratum to the overall adjusted *AR* was 5.8%. Summing the contributions of all nine strata yielded an adjusted *AR* estimate of 70.0%, thus lower than both the crude and Mantel–Haenszel adjusted *AR* estimates. The corresponding standard error estimate was 5.8%, higher than the standard error estimate from the Mantel–Haenszel approach because fewer assumptions were made. Namely, the odds ratio was not assumed common to all strata, so that nine separate odds ratios had to be estimated (one for each stratum) rather than a single common odds ratio from all strata. To circumvent the problem of empty cells, the standard error estimate was obtained after assigning the value 0.5 to all zero cells.

A natural alternative to generalize these approaches is to use adjustment procedures based on regression models, in order to take advantage of their flexible and unified approach to efficient parameter estimation and hypothesis testing. Regression models allow one to take into account adjustment factors as well as interaction of exposures with some or all adjustment factors. This approach was first used by Walter (1976), Sturmans et al. (1977), and Fleiss (1979) followed by

Deubner et al. (1980) and Greenland (1987). The full generality and flexibility of the regression approach was exploited by Bruzzi et al. (1985) who developed a general *AR* estimate based on rewriting *AR* as

$$1 - \sum_{j=1}^{J} \sum_{i=0}^{I} \rho_{ij} RR_{i|j}^{-1}.$$

Quantities ρ_{ij} represent the proportion of diseased individuals with level i of exposure ($i = 0$ at the reference level, $i = 1, \ldots, I$ for exposed levels) and j of confounding and can be estimated from cohort or case-control data (or cross-sectional survey data) using the observed proportions. The quantity $RR_{i|j}^{-1}$ represents the inverse of the rate ratio, risk ratio, or odds ratio depending on the context, for level i of exposure at level j of confounding. It can be estimated from regression models (see Sect. 4.4), both for cohort and case-control data (as well as cross-sectional data), which allows confounding and interactions to be accounted for. Hence, this regression-based approach to *AR* estimation allows control for confounding and interaction and can be used for the main epidemiological designs. Depending on the design, conditional or unconditional logistic, log-linear or Poisson models can be used. Variance estimators were developed based on an extension of the delta method to implicitly related random variables in order to take into account the variability in estimates of terms ρ_{ij} and $RR_{i|j}^{-1}$ as well as their correlations (Basu and Landis 1995; Benichou and Gail 1989, 1990b). This regression approach includes the crude and two stratification approaches as special cases and offers additional options (Benichou 1991). The unadjusted approach corresponds to models for $RR_{i|j}^{-1}$ with exposure only. The Mantel–Haenszel approach corresponds to models with exposure and confounding factors but no interaction terms between exposure and adjustment factors. The weighted-sum approach corresponds to fully saturated models with all interaction terms between exposure and confounding factors. Intermediate models are possible, for instance, models allowing for interaction between exposure and only one confounder, or models in which the main effects of some confounders are not modeled in a saturated way.

Example 4.3. (Continued)
Still considering the more restrictive definition of the reference category for daily alcohol consumption, an unconditional logistic model (see Sect. 4.4.6.3) with two parameters, one general intercept and one parameter for elevated alcohol consumption, was fit, ignoring smoking and age. The resulting unadjusted odds ratio estimate was 5.9 as in Sect. 4.5.1.6. The formula above for $1 - AR$ reduced to a single sum with two terms ($i = 0, 1$) corresponding to unexposed and exposed categories, respectively. The resulting unadjusted *AR* estimate was 70.9% (standard error estimate of 5.1), identical to the crude *AR* estimate in Sect. 4.5.1.6. Adding eight terms for smoking and age in the logistic model increased the fit significantly ($p < 0.001$, likelihood ratio test) and yielded an adjusted odds ratio estimate of 6.3, slightly higher than the Mantel–Haenszel odds ratio estimate of 6.2 (see above). This resulted in an adjusted *AR* estimate of 71.9%, slightly higher than the corresponding Mantel–Haenszel *AR* estimate of 71.6%, and with a slightly lower standard error estimate of 5.0%. Adding two terms for interactions of smoking with alcohol consumption (thus

allowing for different odds ratio estimates depending on smoking level) resulted in a decreased *AR* estimate of 70.3% (with a higher standard error estimate of 5.4% because of the additional parameters estimated). Adding six more terms allowed for all two-by-two interactions between alcohol consumption and joint age × smoking level and yielded a fully saturated model. Thus, nine odds ratios for alcohol consumption were estimated (one for each stratum) as with the weighted-sum approach. This resulted in little change as regards *AR*, with an *AR* estimate of 70.0%, identical to the *AR* estimate with the weighted-sum approach, which precisely corresponds to a fully saturated model. The corresponding standard error estimate was increased to 5.6% due to the estimation of additional parameters.

A modification of Bruzzi et al.'s approach was developed by Greenland and Drescher (1993) in order to obtain full maximum-likelihood estimates of *AR*. The modification consists in estimating the quantities ρ_{ij} from the regression model rather than simply relying on the observed proportions of cases. The two model-based approaches seem to differ very little numerically (Greenland and Drescher 1993). Greenland and Drescher's approach might be more efficient in small samples although no difference was observed in simulations of the case-control design even for samples of 100 cases and 100 controls (Greenland and Drescher 1993). It might be less robust to model misspecification, however, as it relies more heavily on the *RR* or odds ratio model used. Finally, it does not apply to the conditional logistic model, and if that model is to be used (notably, in case-control studies with individual matching), the original approach of Bruzzi et al. is the only possible choice.

Detailed reviews of adjusted *AR* estimation (Benichou 1991, 2001; Coughlin et al. 1994; Gefeller 1992) are available. Alternative methods to obtain estimates of variance and confidence intervals for *AR* have been developed either based on re-sampling techniques (Gefeller 1992; Greenland 1992; Kahn et al. 1998; Kooperberg and Petitti 1991; Llorca and Delgado-Rodriguez 2000; Uter and Pfahlberg 1999) or on quadratic equations (Lui 2001a, b, 2003).

4.5.1.8 Final Notes and Additional References

General problems of *AR* definition, interpretation, and usefulness as well as properties have been reviewed in detail (Benichou 2000b; Gefeller 1992; Miettinen 1974; Rockhill et al. 1998a, b; Walter 1976). Special issues were reviewed by Benichou (2000b, 2001). They include estimation of *AR* for risk factors with multiple levels of exposure or with a continuous form, multiple risk factors, recurrent disease events, and disease classification with more than two categories. They also include assessing the consequences of exposure misclassification on *AR* estimates. Specific software for attributable risk estimation (Kahn et al. 1998; Mezzetti et al. 1996) as well as a simplified approach to confidence interval estimation (Daly 1998) have been developed to facilitate implementation of methods for attributable risk estimation. Finally, much remains to be done to promote proper use and interpretation of *AR* as illustrated in a literature review (Uter and Pfahlberg 2001).

4.5.2 Attributable Risk Among the Exposed

The attributable risk in the exposed (AR_E) or attributable fraction in the exposed is defined as the following ratio (Cole and MacMahon 1971; Levin 1953; MacMahon and Pugh 1970; Miettinen 1974):

$$AR_E = \left\{\Pr(D|E) - \Pr(D|\overline{E})\right\} / \Pr(D|E), \tag{4.16}$$

where $\Pr(D|E)$ is the probability of disease in the exposed individuals (E) and $\Pr(D|\overline{E})$ is the hypothetical probability of disease in the same subjects but with all exposure eliminated. Depending on the context, these probabilities will refer to disease risk or may be replaced with incidence rates (see Sect. 4.5.1). AR_E can be rewritten as

$$AR_E = (RR - 1)/RR, \tag{4.17}$$

where RR denotes the risk or rate ratio. Following Greenland and Robins (1988), Rothman and Greenland (1998, Chap. 4) proposed to use the terms "excess fraction" for the definition of AR_E based on risks or risk ratios and "rate fraction" for the definition of AR_E based on rates or rate ratios.

Like AR, AR_E lies between 0 and 1 when exposures considered are risk factors ($RR > 1$) with a maximal limiting value of 1, is equal to zero in the absence of association between exposure and disease ($RR = 1$), and is negative for protective exposures ($RR < 1$).

As for AR, AR_E has an interpretation as a measure of the disease burden attributable or at least related to one or several exposures among the exposed subjects. Consequently, AR_E could be used to assess the potential impact of prevention programs aimed at eliminating exposure from the population. These interpretations are subject to the same limitations as corresponding interpretations for AR, however (see Sect. 4.5.1.4). Moreover, AR_E does not have a clear public health interpretation because it does not depend on the exposure prevalence but only on the risk or rate ratio of which it is merely a one-to-one transformation. For the assessment of the relative impact of several exposures, AR_E will not be an appropriate measure since AR_E for different exposures refers to different groups of subjects in the population (i.e., subjects exposed to each given exposure).

AR_E being a one-to-one function of RR, issues of estimability, and estimation for AR_E are similar to those for RR. They depend on whether rates or risks are considered. For case-control studies, odds ratios can be used. Greenland (1987) specifically derived adjusted point estimates and confidence intervals for AR_E based on the Mantel–Haenszel approach.

4.5.3 Sequential and Partial Attributable Risks

Upon considering multiple exposures, separate ARs can be estimated for each exposure as well as the overall AR for all exposures jointly. Except in very special circumstances worked out by Walter (1983) (i.e., lack of joint exposure or additive effects of exposures on disease risk or rate), the sum of separate *AR* estimates over all exposures considered will not equal the overall *AR* estimate.

Because this property is somewhat counterintuitive and generates misinterpretations, three alternative approaches have been suggested: one based on considering variance decomposition methods (Begg et al. 1998) rather than estimating *AR*, one based on estimating assigned share or probability of causation of a given exposure with relevance in litigation procedures for individuals with multiple exposures (Cox 1984, 1985; Lagakos and Mosteller 1986; Seiler 1987; Seiler and Scott 1987; Benichou 1993b; McElduff et al. 2002), and one based on an extension of the concept of *AR* (Eide and Gefeller 1995; Land et al. 2001). This last approach relies on partitioning techniques (Gefeller et al. 1998; Land and Gefeller 1997) and keeps with the framework of *AR* estimation by introducing the sequential *AR* that generalizes the concept of *AR*. The principle is to define an order among the exposures considered. Then, the contribution of each exposure is assessed sequentially according to that order. The contribution of the first exposure considered is calculated as the standard *AR* for that exposure separately. The contribution of the second exposure is obtained as the difference between the joint *AR* estimate for the first two exposures and the separate *AR* estimate for the first exposure, the contribution of the third exposure is obtained as the difference between the joint *AR* estimates for the first three and first two exposures, etc. Thus, a multidimensional vector consisting of contributions of each separate exposure is obtained.

These contributions are meaningful in terms of potential prevention programs that consider successive rather than simultaneous elimination of exposures from the population. Indeed, each step yields the additional contribution of the elimination of a given exposure once higher-ranked exposures are eliminated. At some point, additional contributions may become very small, indicating that there is not much point in considering extra steps. By construction, these contributions sum to the overall *AR* for all exposures jointly, which constitutes an appealing property. Of course, separate vectors of contributions are obtained for different orders. Meaningful orders depend on practical possibilities in implementing potential prevention programs in a given population. Average contributions can be calculated for each given exposure by calculating the mean of contributions corresponding to that exposure over all possible orders. These average contributions have been termed partial attributable risks (Eide and Gefeller 1995) and represent another potentially useful measure. Methods for visualizing sequential and partial *AR*s are provided by Eide and Heuch (2001). An illustration is given by Fig. 4.1. A detailed review of properties, interpretation, and variants of sequential and partial *AR*s was provided by Land et al. (2001).

Example 4.3. (Continued)
Smoking is also a known risk factor of esophageal cancer, so it is important to estimate the impact of smoking and the joint impact of smoking and alcohol consumption on esophageal cancer in addition to that of alcohol consumption alone. Using the first category (i.e., 0–9 g/day) as the reference level of smoking, there were 78 cases in the reference level of smoking, 122 cases in the exposed level (i.e., 10+ g/day), 447 controls in the reference level, and 328 controls in the exposed level. From these data, the crude odds ratio estimate for smoking at least 10 g/day was 2.1, and the crude *AR* estimate for smoking at least 10 g/day was 32.4%. Moreover, there were 9 cases and 252 controls in the joint

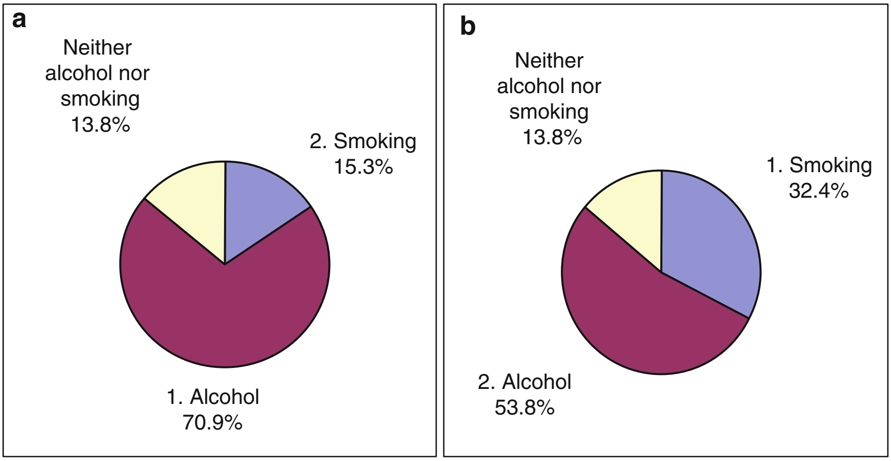


Fig. 4.1 Sequential attributable risk estimates for elevated alcohol consumption (80+ g/day) and heavy smoking (10+ g/day) for two different orders of removal (*left panel* (**a**): alcohol, then smoking; *right panel* (**b**): smoking, then alcohol) – case-control data on esophageal cancer (Tuyns et al. 1977; cf. Example 4.3)

reference level of alcohol consumption and smoking (i.e., 0–39 g/day of alcohol and 0–9 g/day of tobacco), which yielded a crude joint odds ratio estimate of 10.2 and a crude joint *AR* estimate for drinking at least 40 g/day of alcohol or smoking at least 10 g/day of tobacco of 86.2%.

Furthermore, the crude *AR* estimate for alcohol consumption of at least 40 g/day was estimated at 70.9% in Sect. 4.5.1.6. Hence, considering the first order of risk factor removal (i.e., eliminating alcohol consumption above 39 g/day followed by eliminating smoking above 9 g/day) yields sequential *AR* estimates of 70.9% for elevated daily alcohol consumption and 86.2% − 70.9% = 15.3 percentage points for heavy smoking so that, once elevated alcohol consumption is eliminated, the additional impact of eliminating heavy smoking appears rather limited (Fig. 4.1a). Considering the second order (i.e., eliminating heavy smoking first) yields sequential *AR* estimates of 32.4% for heavy smoking and 86.2% − 32.4% = 53.8 percentage points for elevated alcohol consumption so that, once heavy smoking is eliminated, the additional impact of eliminating elevated alcohol consumption remains major (Fig. 4.1b). A summary of these results is provided by partial *AR*s for elevated alcohol consumption and heavy smoking, with estimated values of 62.4% and 23.9%, respectively, again reflecting the higher impact of elevated alcohol consumption on esophageal cancer.

4.5.4 Preventable and Prevented Fractions

When considering a protective exposure or intervention, an appropriate alternative to *AR* is the preventable or prevented fraction (*PF*) defined as the ratio (Miettinen 1974):

$$PF = \left\{\Pr(D|\overline{E}) - \Pr(D)\right\} / \Pr(D|\overline{E}),$$

where $\Pr(D)$ is the probability of disease in the population, which may have some exposed (E) and some unexposed ($\overline{E}$) individuals, and $\Pr(D|\overline{E})$ is the hypothetical probability of disease in the same population but with all (protective) exposure eliminated. Depending on the context, these probabilities will refer to disease risk or may be replaced with incidence rates (see sections above). *PF* can be rewritten as

$$PF = p_E(1 - RR), \tag{4.19}$$

a function of both the prevalence of exposure, p_E, and the risk or rate ratio, *RR*. Thus, a strong association between exposure and disease may correspond to a high or low value of *PF* depending on the prevalence of exposure, as for *AR*. Moreover, portability is not a typical property of *PF*, as for *AR*. As for *AR* again, it may be useful to compare *PF* estimates among population subgroups to target prevention efforts to specific subgroups with a potentially high impact (as measured by the *PF*).

For a protective factor ($RR < 1$), *PF* lies between 0 and 1 and increases with the prevalence of exposure and the strength of the association between exposure and disease.

PF measures the impact of an association between a protective exposure and disease at the population level. It has a public health interpretation as the proportion of disease cases averted ("prevented fraction") in relation to the presence of a protective exposure or intervention in the population, among the totality of cases that would have developed in the absence of that factor or intervention in the population. In this case, it is useful to assess prevention programs a posteriori. Alternatively, it can be used to assess prevention programs a priori by measuring the proportion of cases that could be potentially averted ("preventable fraction") if a protective exposure or intervention was introduced de novo in the population (Gargiullo et al. 1995). These interpretations are subject to the same limitations as corresponding interpretations for *AR*, however (see Sect. 4.5.1.4).

PF and *AR* are mathematically related through (Walter 1976)

$$1 - PF = 1/(1 - AR). \tag{4.20}$$

From Eq. 4.20, it appears that, for a protective factor, *PF* estimates will usually differ from *AR* estimates obtained by reversing the coding of exposure. This follows from the respective definitions of *AR* and *PF*. While *AR*, with reverse coding, measures the potential reduction in disease occurrence that could be achieved if all subjects in the current population became exposed, *PF* measures the reduction in disease occurrence obtained from introducing exposure at the current prevalence in a formally unexposed population (Benichou 2000c).

In view of Eq. 4.20, estimability and estimation issues are similar for *AR* and *PF*. Specific *PF* adjusted point and confidence interval estimates were derived using the Mantel–Haenszel approach (Greenland 1987) and weighted-sum approaches (Gargiullo et al. 1995).

4.5.5 Generalized Impact Fraction

The generalized impact fraction (*GIF*) or generalized attributable fraction was introduced by Walter (1980) and Morgenstern and Bursic (1982) as the ratio

$$GIF = \{\Pr(D) - \Pr^*(D)\}/\Pr(D), \tag{4.21}$$

where $\Pr(D)$ and $\Pr^*(D)$, respectively, denote the probability of disease under the current distribution of exposure and under a modified distribution of exposure. As for *AR* and *PF*, these probabilities denote risks or can be replaced by incidence rates depending on the context.

The generalized impact fraction not only depends on the association between exposure and disease as well as the current distribution (rather than just the prevalence) of exposure but also on the target distribution of exposure considered that will yield $\Pr^*(D)$. It is a general measure of impact that includes *AR* and *PF* as special cases. *AR* contrasts the current distribution of exposure with a modified distribution defined by the absence of exposure. Conversely, *PF* contrasts a distribution defined by the absence of exposure with the current distribution of exposure (prevented fraction) or target distribution of exposure (preventable fraction).

The generalized impact fraction can be interpreted as the fractional reduction of disease occurrence that would result from changing the current distribution of exposure in the population to some modified distribution. Thus, it can be used to assess prevention programs or interventions, targeting all subjects or subjects at specified levels, and aimed at modifying or shifting the exposure distribution (reducing exposure) but not necessarily eliminating exposure. For instance, heavy smokers could be specifically targeted by interventions rather than all smokers. The special *AR* case corresponds to the complete elimination of exposure by considering a modified distribution putting unit mass on the lowest risk configuration and can be used to assess interventions aimed at eliminating (rather than reducing) exposure. Alternatively, the general impact fraction could be used to assess the increase in disease occurrence as a result of exposure changes in the population, such as the increase in breast cancer incidence as a result of delayed childbearing (Kleinbaum et al. 1982, Chap. 9). Such interpretations are subject to the same limitations as for *AR* and *PF* (see Sect. 4.5.1.4).

The generalized impact fraction has been used, for instance, by Lubin and Boice (1989) who considered the impact on lung cancer of a modification in the distribution of radon exposure consisting in truncating the current distribution at various thresholds and by Wahrendorf (1987) who examined the impact of various changes in dietary habits on colorectal and stomach cancers.

Issues of estimability are similar to those for *AR* and *PF*. Methods to estimate the generalized impact fraction are similar to methods for estimating *AR* and *PF*. However, unlike for *AR* or *PF*, it might be useful to retain the continuous nature of exposures to define the modification of the distribution considered (for instance, a shift in the distribution), and extensions of methods for estimating *AR* for continuous factors (Benichou and Gail 1990b) are relevant in this context. Drescher and Becher (1997) proposed extending model-based approaches of Bruzzi et al. (1985) and Greenland and Drescher (1993) to estimate the generalized impact fraction in case-control studies and considered continuous as well as categorical exposures.

4.5.6 Person-Years of Life Lost

Person-years of life lost (or potential years of life lost, PYLL) for a given cause of death is a measure defined as the difference between current life expectancy of the population and potential life expectancy with the cause of death eliminated (Smith 1998). For instance, one may be interested in PYLL due to prostate cancer in men, breast cancer in women, or cancer as a whole (all sites) in men and women. Methods for estimating PYLL rely on calculating cause-deleted life tables. Total PYLL at the population level or average PYLL per person may be estimated. As an example, a recent report from the Surveillance, Epidemiology, and End Results (SEER) estimated that 8.4 million years of life overall were lost due to cancer in the US population (both sexes, all races) in the year 2001, with an average value of potential years of life lost per person of 15.1 years. Corresponding numbers were 779,900 years overall and 18.8 years on average for breast cancer in women and 275,200 years overall and 9.0 years on average for prostate cancer in men (Ries et al. 2004).

PYLL represents an assessment of the impact of a given disease. Thus, it is not directly interpretable as a measure of exposure impact, except perhaps for diseases with a dominating risk factor, such as asbestos exposure for mesothelioma or human papilloma virus for cervical cancer.

However, it is possible to obtain a corresponding measure of the impact of a given exposure by converting PYLL due to a particular cause of death to PYLL due to a particular exposure. Estimation of an exposure-specific PYLL is obtained through applying an *AR* estimate for that exposure to the disease-specific PYLL, namely, calculating the product PYLL times *AR*, which yields the fraction of PYLL attributable to exposure. In this process, several causes of deaths may have to be considered. For instance, the fractions of PYLL for mesothelioma and lung cancer would need to be added in order to obtain the overall PYLL for asbestos exposure. In contrast with *AR* that provides a measure of exposure impact as a fraction of disease incidence (or death), such calculations of PYLL will provide a measure of exposure impact on the life expectancy scale. As for *AR*, the impact of a given exposure on the PYLL scale will depend on the prevalence of exposure in the population and strength of association between exposure and disease(s). Moreover, it will depend critically on the age distribution of exposure-associated diseases and their severity, i.e. case fatality.

4.6 Other Topics

4.6.1 Standardization of Risks and Rates

Risks and rates can usually not be directly compared between countries, regions, or time periods because of differences in age structure. For example, an older population may appear to have higher rates of certain cancers, not because of the presence of risk factors but because of the higher age itself. This is a form of confounding. In the tradition of demography, so-called standardization is applied to reported rates and risks to adjust for differences in age and possibly other confounders. Direct standardization is the most commonly used technique. It proceeds by forming a weighted average of age-specific rates or risks, where the weights reflect a known population structure. This structure is typically chosen as that of a country in a given census year, the so-called standard population. A directly standardized rate can be written:

$$SR = \Sigma n_j^S h_j / \Sigma n_j^S = \Sigma w_j^S h_j, \tag{4.22}$$

where n_j^S is the number of individuals in age group j in the standard population, h_j are age specific rates in the population under study, and w_j^S are weights such that $\Sigma w_j^S = 1$. A standardized risk can be computed in the same manner. Since the weights are fixed and not estimated, the variance of the estimated standardized rate is

$$\text{var}(\Sigma w_j^S h_j) = \Sigma (w_j^S)^2 h_j, \tag{4.23}$$

based on the Poisson assumption for the age specific rates. For risks, the binomial assumption may be used for the age-specific risks.

When age-specific rates or risks are not available in the exposed population under study, so-called indirect standardization may be used. This technique is less common but requires knowledge only of the age distribution, and not the age-specific rates, in the population under study. The approach is particularly useful if the exposed cohort is relatively small and the age-specific incidence or mortality rates thus are unstable. The term is somewhat of a misnomer, as the technique in effect employs direct standardization of the rates in the unexposed (reference) to the exposed population where the age distribution of the exposed group is the standard. The *indirect* step serves only to project back to a rate standardized to the unexposed reference population. The indirectly standardized rate is obtained by $(SMR)(CR^O)$, where *SMR* is the *standardized mortality* or *morbidity ratio* (see below) and CR^O is the *crude* (i.e., original overall) rate in a reference population that provides stratum-specific rates.

The standardized mortality or morbidity ratio is a ratio between observed and expected event counts, where the expected count is based on age-specific rates or risks in a reference population, which is a non-exposed or general population group. Then, the standardized mortality (or morbidity or incidence) ratio

$$SMR\ or\ SIR = D_E/E_0 = \Sigma n_j h_j / \Sigma n_j h_{0j},$$

where D_E is the number of events in the exposed and E_0 is the expected number of events obtained from the rates h_{0j} in the unexposed applied to the sample composition of the exposed. The *SMR* can also be rewritten as a weighted average of sex- and age-specific (say) rate ratios h_j/h_{0j} with weights $w_j = n_j h_{0j}$. It can be shown that these weights minimize the standard error of the weighted average (Breslow and Day 1987, Chap. 2) as long as the rates in the reference population are assumed to be known rather than estimated. Stratified analyses as discussed above, on the other hand, choose weights that minimize the standard errors when the rates are estimated among both the exposed and the unexposed. The *SMR* has the advantage that age- and sex-specific rates are not needed for the exposed group.

The denominator of the *SMR* is generally obtained from age- and sex-specific rates in the entire regional population. This allows the random variation of the denominator to be considered to be none, and confidence intervals can be based on the estimate $1/D_E^{1/2}$ for the standard error of ln(*SMR*). This standard error computation also assumes that the events among the exposed are uncorrelated (do not cluster) or, more specifically, that the event count follows a Poisson distribution.

It may be noted that directly standardized rates are based on a choice of standard population to generate weights. While the weights used for the *SMR* result from the composition of the comparison group and do not involve a true standard population, the weights used in direct standardization are external as they result from information outside the samples being compared. In principle, the latter weights are similar to survey weights applied in, e.g., the National Health and Nutrition Examination Survey (NHANES), where the sample must be standardized to the US population to account for the methodology used in drawing it. While improving external validity, weights from direct standardization and survey weights always result in loss of statistical efficiency, i.e., standard errors will be larger than for crude or non-weighted rates and risks. In contrast, many of the methods to adjust for confounding discussed in Sect. 4.4 are internal to the specific comparison and designed to optimize statistical efficiency.

4.6.2 Measures Based on Prevalence

Prevalence is the number of cases either at a given point in time (point prevalence) or over a time period (period prevalence) divided by the population size. Prevalence can be easier to obtain than incidence. For example, a population survey can determine how many individuals in a population suffer from a given illness or health condition at a point in time.

Measures of association based on prevalence parallel those for risk (for point prevalence) or incidence rates (for period prevalence). For example, one can form prevalence ratios, prevalence differences, and prevalence odds ratios. Measures of impact based on prevalence can also be obtained.

Prevalence and the measures of association based on it are useful entities for health policy planning and for determining the level of services needed for individuals with a given health condition in the population. It is usually considered less useful for studying the etiology of a disease. The reason for this is that under certain assumptions, prevalence of a disease equals its incidence multiplied by its duration (Kleinbaum et al. 1982, Chap. 8). These assumptions are that the population is stable and that both the incidence and prevalence remain constant. Under more general conditions, prevalence still reflects both incidence and duration but in a more complex manner. For a potentially fatal or incurable disease, duration means survival, and the exposures that increase incidence may reduce or increase survival and hence the association of an exposure with prevalence may be very different than its association with incidence. On the other hand, when a disease or condition can be of limited duration due to recovery or cure, and its duration is maintained by the same exposures that caused it, prevalence can be more meaningful than incidence. For example, it is conceivable that weight gain in a person may have caused hypertension, and when the person loses the same amount of weight, she/he moves out of being hypertensive. In this latter case, the prevalence ratio between the percentages with hypertension in those exposed and unexposed to the risk factor captures the increase in the risk of living with the condition caused by the exposure, while the incidence ratio captures only part of the etiological association.

4.7 Conclusions

Disease frequency is measured through the computation of incidence rates or estimation of disease risk. Both measures are directly accessible from cohort data. They can be obtained from case-control data only if they are complemented by follow-up or population data. Using regression techniques, methods are available to derive incidence rates or risk estimates specific to a given exposure profile. Exposure-specific risk estimates are useful in individual prediction.

A wide variety of options and techniques are available for measuring association. The odds ratio is presently the most often used measure of association for both cohort and case-control studies. Adjustment for confounding is key in all analyses of observational studies and can be pursued by standardization, stratification, and by regression techniques. The flexibility of the latter, especially in the generalized linear model framework, and availability of computer software, has made it widely applied in the last several years.

Several measures are available to assess the impact of an exposure in terms of the occurrence of new disease cases at the population level, among which the attributable risk is the most commonly used. Several approaches have been developed to derive adjusted estimates of the attributable risk from case-control as well as cohort data, either based on stratification or on more flexible regression techniques. The concept of attributable risk has been extended to handle preventive exposures, multiple exposures, as well as assessing the impact of various modifications of the exposure distribution rather than the mere elimination of exposure.

References

Aalen O (1978) Nonparametric estimation of partial transition probabilities in multiple decrement models. Ann Stat 6:534–545

Aalen O, Johansen S (1978) An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. Scand J Stat 5:141–150

American Cancer Society (1992) Cancer facts and figures. American Cancer Society, Atlanta

Ames BN, Gold LS, Willett WC (1995) The causes and prevention of cancer. Proc Natl Acad Sci U S A 254:1131–1138

Andersen PK, Gill RD (1982) Cox's regression models for counting processes: a large-sample study. Ann Stat 4:1100–1120

Anderson KM, Wilson PW, Odell PM, Kannel WB (1991) Cardiovascular disease risk profiles. A statement for health professionals. Circulation 83:356–362

Anderson SJ, Ahnn S, Duff K (1992) NSABP breast cancer prevention trial risk assessment program, version 2. Department of Biostatistics, University of Pittsburgh, Pittsburgh

Basu S, Landis JR (1995) Model-based estimation of population attributable risk under cross-sectional sampling. Am J Epidemiol 142:1338–1343

Begg CB (2001) The search for cancer risk factors: when can we stop looking? Am J Public Health 91:360–364

Begg CB, Satagopan JM, Berwick M (1998) A new strategy for evaluating the impact of epidemiologic risk factors for cancer with applications to melanoma. J Am Stat Assoc 93: 415–426

Benichou J (1991) Methods of adjustment for estimating the attributable risk in case-control studies: a review. Stat Med 10:1753–1773

Benichou J (1993a) A computer program for estimating individualized probabilities of breast cancer. Comput Biomed Res 26:373–382

Benichou J (1993b) Re: "Methods of adjustment for estimating the attributable risk in case-control studies: a review" (letter). Stat Med 12:94–96

Benichou J (2000a) Absolute risk. In: Gail MH, Benichou J (eds) Encyclopedia of epidemiologic methods. Wiley, Chichester, pp 1–17

Benichou J (2000b) Attributable risk. In: Gail MH, Benichou J (eds) Encyclopedia of epidemiologic methods. Wiley, Chichester, pp 50–63

Benichou J (2000c) Preventable fraction. In: Gail MH, Benichou J (eds) Encyclopedia of epidemiologic methods. Wiley, Chichester, pp 736–737

Benichou J (2001) A review of adjusted estimators of the attributable risk. Stat Methods Med Res 10:195–216

Benichou J, Gail MH (1989) A delta-method for implicitely defined random variables. Am Stat 43:41–44

Benichou J, Gail MH (1990a) Estimates of absolute cause-specific risk in cohort studies. Biometrics 46:813–826

Benichou J, Gail MH (1990b) Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models. Biometrics 46:991–1003

Benichou J, Gail MH (1995) Methods of inference for estimates of absolute risk derived from population-based case-control studies. Biometrics 51:182–194

Benichou J, Wacholder S (1994) A comparison of three approaches to estimate exposure-specific incidence rates from population-based case-control data. Stat Med 13:651–661

Benichou J, Gail MH, Mulvihill JJ (1996) Graphs to estimate an individualized risk of breast cancer. J Clin Oncol 14:103–110

Benichou J, Byrne C, Gail MH (1997) An approach to estimating exposure-specific rates of breast cancer from a two-stage case-control study within a cohort. Stat Med 16:133–151

Berkson J (1958) Smoking and lung cancer. Some observations on two recent reports. J Am Stat Assoc 53:28–38

Birch MW (1964) The detection of partial associations, I: the 2×2 case. J R Stat Soc B 27:313–324

Borgan Ø (1998) Nelson-Aalen estimator. In: Armitage P, Colton T (eds) Encyclopedia of epidemiologic methods. Wiley, Chichester, pp 2967–2972

Breslow NE (1981) Odds ratio estimators when the data are sparse. Biometrika 68:73–84

Breslow NE, Day NE (1980) Statistical methods in cancer research vol 1: the analysis of case-control studies. International Agency for Research on Cancer Scientific Publications No. 32, Lyon

Breslow NE, Day NE (1987) Statistical methods in cancer research vol II: the design and analysis of cohort studies. International Agency for Research on Cancer Scientific Publications No. 82, Lyon

Breslow NE, Lubin JH, Marek P, Langholz B (1983) Multiplicative models and cohort analysis. J Am Stat Assoc 78:1–12

Bruzzi P, Green SB, Byar DP, Brinton LA, Schairer C (1985) Estimating the population attributable risk for multiple risk factors using case-control data. Am J Epidemiol 122:904–914

Chiang CL (1968) Introduction to stochastic processes in biostatistics. Wiley, New York

Colditz G, DeJong W, Hunter D, Trichopoulos D, Willett W (eds) (1996) Harvard report on cancer prevention, vol 1. Cancer Causes Control 7(suppl):S3–S59

Colditz G, DeJong W, Hunter D, Trichopoulos D, Willett W (eds) (1997) Harvard report on cancer prevention, vol 2. Cancer Causes Control 8(suppl):S1–S50

Cole P, MacMahon B (1971) Attributable risk percent in case-control studies. Br J Prev Soc Med 25:242–244

Cornfield J (1951) A method for estimating comparative rates from clinical data: applications to cancer of the lung, breast and cervix. J Natl Cancer Inst 11:1269–1275

Cornfield J (1956) A statistical problem arising from retrospective studies. In: Neyman J (ed) Proceedings of the third Berkeley symposium, vol IV. University of California Press, Monterey, pp 133–148

Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EI (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. J Natl Cancer Inst 22: 173–203

Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, Wieand HS (1999) Validation studies for models projecting the risk of invasive and total breast cancer incidence. J Natl Cancer Inst 91:1541–1548

Coughlin SS, Benichou J, Weed DL (1994) Attributable risk estimation in case-control studies. Epidemiol Rev 16:51–64

Cox DR (1972) Regression models and lifetables (with discussion). J R Stat Soc B 34:187–220

Cox DR (1975) Partial likelihood. Biometrika 62:269–276

Cox LA (1984) Probability of causation and the attributable proportion of risk. Risk Anal 4:221–230

Cox LA (1985) A new measure of attributable risk for public health applications. Manage Sci 7:800–813

Cutler SJ, Ederer F (1958) Maximum utilization of the life table method in analyzing survival. J Chronic Dis 8:699–712

Daly LE (1998) Confidence limits made easy: interval estimation using a substitution method. Am J Epidemiol 147:783–790

Deubner DC, Wilkinson WE, Helms MJ, Tyroler HA, Hames CG (1980) Logistic model estimation of death attributable to risk factors for cardiovascular disease in Evans County, Georgia. Am J Epidemiol 112:135–143

Doll R, Peto R (1981) The causes of cancer. Oxford University Press, New York

Dorey FJ, Korn EL (1987) Effective sample sizes for confidence intervals for survival probabilities. Stat Med 6:679–687

Drescher K, Becher H (1997) Estimating the generalized attributable fraction from case-control data. Biometrics 53:1170–1176

Dupont DW (1989) Converting relative risks to absolute risks: a graphical approach. Stat Med 8:641–651

Easton DF, Peto J, Babiker AG (1991) Floating absolute risk: an alternative to relative risk in survival and case-control analysis avoiding an arbitrary reference group. Stat Med 10:1025–1035

Eide GE, Gefeller O (1995) Sequential and average attributable fractions as aids in the selection of preventive strategies. J Clin Epidemiol 48:645–655

Eide GE, Heuch I (2001) Attributable fractions: fundamental concepts and their vizualization. Stat Methods Med Res 10:159–193

Ejigou A (1979) Estimation of attributable risk in the presence of confounding. Biom J 21:155–165

Elandt-Johnson RC (1977) Various estimators of conditional probabilities of death in follow-up studies. Summary of results. J Chronic Dis 30:247–256

Elveback L (1958) Estimation of survivorship in chronic disease: the "actuarial" method. J Am Stat Assoc 53:420–440

Fleiss JL (1979) Inference about population attributable risk from cross-sectional studies. Am J Epidemiol 110:103–104

Fleiss JL, Dunner DL, Stallone F, Fieve RR (1976) The life table: a method for analyzing longitudinal studies. Arch Gen Psychiatry 33:107–112

Fisher B, Costantino JP, Wickerham L, Redmond CK, Kavanah M, Cronin WM, Vogel V, Robidoux A, Dimitrov N, Atkins J, Daly M, Wieand S, Tan-Chiu E, Ford L, Womark N, other National Surgical Adjuvant Breast and Bowel project Investigators (1998) Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. J Natl Cancer Inst 90:1371–1388

Gail MH (1975) Measuring the benefit of reduced exposure to environmental carcinogens. J Chronic Dis 28:135–147

Gail MH, Benichou J (1994) Validation studies on a model for breast cancer risk (editorial). J Natl Cancer Inst 86:573–575

Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 81:1879–1886

Gail MH, Costantino JP, Bruant J, Croyle R, Freedman L, Helzsouer K, Vogel V (1999) Weighing the risks and benefits of Tamoxifen treatment for preventing breast cancer. J Natl Cancer Inst 91:1829–1846

Gargiullo PM, Rothenberg R, Wilson HG (1995) Confidence intervals, hypothesis tests, and sample sizes for the prevented fraction in cross-sectional studies. Stat Med 14:51–72

Gefeller O (1990) Theory and application of attributable risk estimation in cross-sectional studies. Stat Appl 2:323–331

Gefeller O (1992) The bootstrap method for standard errors and confidence intervals of the adjusted attributable risk [letter]. Epidemiology 3:271–272

Gefeller O (1995) Definitions of attributable risk-revisited. Public Health Rev 23:343–355

Gefeller O, Land M, Eide GE (1998) Averaging attributable fractions in the multifactorial situation: assumptions and interpretation. J Clin Epidemiol 51:437–451

Gray RJ (1988) A class of k-sample tests for comparing the cumulative incidence of a competing risk. Ann Stat 16:1141–1151

Greenland S (1981) Multivariate estimation of exposure-specific incidence from case-control studies. J Chronic Dis 34:445–453

Greenland S (1984) Bias in methods for deriving standardized mortality ratio and attributable fraction estimates. Stat Med 3:131–141

Greenland S (1987) Variance estimators for attributable fraction estimates, consistent in both large strata and sparse data. Stat Med 6:701–708

Greenland S (1992) The bootstrap method for standard errors and confidence intervals of the adjusted attributable risk [letter]. Epidemiology 3:271

Greenland S (2001) Attributable fractions: bias from broad definition of exposure. Epidemiology 12:518–520

Greenland S, Drescher K (1993) Maximum-likelihood estimation of the attributable fraction from logistic models. Biometrics 49:865–872

Greenland S, Morgenstern H (1983) Morgenstern corrects a conceptual error [letter]. Am J Public Health 73:703–704

Greenland S, Robins JM (1988) Conceptual problems in the definition and interpretation of attributable fractions. Am J Epidemiol 128:1185–1197

Greenland S, Thomas DC (1982) On the need for the rare disease assumption. Am J Epidemiol 116:547–553

Greenland S, Schlesselman JJ, Criqui MH (1986) The fallacy of employing standardized regression coefficients and correlations as measures of effect. Am J Epidemiol 123:203–208

Hanley JA, Negassa A, Edwardes MD, Forrester JE (2003) Statistical analysis of correlated data using generalized estimating equations: an orientation. Am J Epidemiol 157:364–375

Hartman LC, Sellers TA, Schaid DJ, Franks TS, Soderberg CL, Sitta DL, Frost MH, Grant CS, Donohue JH, Woods JE, McDonnell SK, Vockley CW, Deffenbaugh A, Couch FJ, Jenkins RB (2001) Efficacy of bilateral prophylactic mastectomy in BRCA1 and BRCA2 gene mutation carriers. J Natl Cancer Inst 93:1633–1637

Henderson BE, Ross RK, Pike MC (1991) Toward the primary prevention of cancer. Science 254:1131–1138

Holford TR (1980) The analysis of rates and of survivorship using log-linear models. Biometrics 36:299–305

Hosmer D, Lemeshow S (1999) Applied survival analysis: regression modeling of time to event data. Wiley, Hoboken

Hoskins KF, Stopfer JE, Calzone K, Merajver SD, Rebbeck TR, Garber JE, Weber BL (1995) Assessment and counseling for women with a family history of breast cancer. A guide for clinicians. J Am Med Assoc 273:577–585

Kahn HA, Sempos CT (1989) Statistical methods in epidemiology. Monographs in epidemiology and biostatistics, vol 12. Oxford University Press, Oxford/New York

Kahn MJ, O'Fallon WM, Sicks JD (1998) Generalized population attributable risk estimation. Technical Report #54, Mayo Foundation, Rochester

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481

Katz D, Baptista J, Azen SP, Pike MC (1978) Obtaining confidence intervals for the risk ratio in a cohort study. Biometrics 34:469–474

Kay R, Schumacher M (1983) Unbiased assessment of treatment effects on disease recurrence and survival in clinical trials. Stat Med 2:41–58

Keiding N, Andersen PK (1989) Nonparametric estimation of transition intensities and transition probabilities: a case study of a two-state Markov process. Appl Stat 38:319–329

Kleinbaum DG, Kupper LL, Morgenstern H (1982) Epidemiologic research: principles and quantitative methods. Lifetime Learning Publications, Belmont

Kooperberg C, Petitti DB (1991) Using logistic regression to estimate the adjusted attributable risk of low birthweight in an unmatched case-control study. Epidemiology 2:363–366

Korn EL, Dorey FJ (1992) Applications of crude incidence curves. Stat Med 11:813–829

Kuritz SJ, Landis JR (1987) Attributable risk estimation from matched-pairs case-control data. Am J Epidemiol 125:324–328

Kuritz SJ, Landis JR (1988a) Summary attributable risk estimation from unmatched case-control data. Stat Med 7:507–517

Kuritz SJ, Landis JR (1988b) Attributable risk estimation from matched case-control data. Biometrics 44:355–367

Lagakos SW, Mosteller F (1986) Assigned shares in compensation for radiation-related cancers (with discussion). Risk Anal 6:345–380

Laird N, Oliver D (1981) Covariance analysis of censored survival data using log-linear analysis techniques. J Am Stat Assoc 76:231–240

Land M, Gefeller O (1997) A game-theoretic approach to partitioning attributable risks in epidemiology. Biom J 39:777–792

Land M, Vogel C, Gefeller O (2001) Partitioning methods for multifactorial risk attribution. Stat Methods Med Res 10:217–230

Landis JR, Heyman ER, Koch GG (1978) Average partial association in three-way contingency tables: a review and discussion of alternative tests. Int Stat Rev 46:237–254

Landis JR, Sharp TJ, Kuritz SJ, Koch G (2000) Mantel-Haenszel methods. In: Gail MH, Benichou J (eds) Encyclopedia of epidemiologic methods. Wiley, Chichester, pp 499–512

Langholz B, Borgan Ø (1997) Estimation of absolute risk from nested case-control data. Biometrics 53:767–774

Last JM (1983) A dictionary of epidemiology. Oxford University Press, New York

Leung HM, Kupper LL (1981) Comparison of confidence intervals for attributable risk. Biometrics 37:293–302

Levin ML (1953) The occurrence of lung cancer in man. Acta Unio Internationalis contra Cancrum 9:531–541

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22

Liddell JC, McDonald JC, Thomas DC (1977) Methods of cohort analysis: appraisal by application to asbestos mining (with discussion). J R Stat Soc A 140:469–491

Lin DY, Psaty BM, Kronmal RA (1998) Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics 54:948–963

Littell AS (1952) Estimation of the t-year survival rate from follow-up studies over a limited period of time. Hum Biol 24:87–116

Llorca J, Delgado-Rodriguez M (2000) A comparison of several procedures to estimate the confidence interval for attributable risk in case-control studies. Stat Med 19:1089–1099

Lubin JH, Boice JD Jr (1989) Estimating Rn-induced lung cancer in the United States. Health Phys 57:417–427

Lui KJ (2001a) Interval estimation of the attributable risk in case-control studies with matched pairs. J Epidemiol Community Health 55:885–890

Lui KJ (2001b) Notes on interval estimation of the attributable risk in cross-sectional sampling. Stat Med 20:1797–1809

Lui KJ (2003) Interval estimation of the attributable risk for multiple exposure levels in case-control studies with confounders. Stat Med 22:2443–2557

Lynch HT, Lynch JF, Rubinstein WS (2001) Prophylactic mastectomy: obstacles and benefits (editorial). J Natl Cancer Inst 93:1586–1587

MacMahon B (1962) Prenatal X-ray exposure and childhood cancer. J Natl Cancer Inst 28:1173–1191

MacMahon B, Pugh TF (1970) Epidemiology: principles and methods. Little, Brown and Co, Boston

Madigan MP, Ziegler RG, Benichou J, Byrne C, Hoover RN (1995) Proportion of breast cancer cases in the United States explained by well-established risk factors. J Natl Cancer Inst 87:1681–1685

Mantel N (1973) Synthetic retrospective studies and related topics. Biometrics 29:479–486

Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 22:719–748

Markush RE (1977) Levin's attributable risk statistic for analytic studies and vital statistics. Am J Epidemiol 105:401–406

Matthews DE (1988) Likelihood-based confidence intervals for functions of many parameters. Biometrika 75:139–144

Mausner JS, Bahn AK (1974) Epidemiology: an introductory text. Saunders, Philadelphia

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. CRC, Boca Raton

McElduff P, Attia J, Ewald B, Cockburn J, Heller R (2002) Estimating the contribution of individual risk factors to disease in a person with more than one risk factor. J Clin Epidemiol 55:588–592

Mehta CR, Patel R, Senchaudhuri P (2000) Efficient Monte Carlo methods for conditional logistic regression. J Am Stat Assoc 95:99–108

Mezzetti M, Ferraroni M, Decarli A, La Vecchia C, Benichou J (1996) Software for attributable risk and confidence interval estimation in case-control studies. Comput Biomed Res 29:63–75

Miettinen OS (1972) Components of the crude risk ratio. Am J Epidemiol 96:168–172

Miettinen OS (1974) Proportion of disease caused or prevented by a given exposure, trait or intervention. Am J Epidemiol 99:325–332

Miettinen OS (1976) Estimability and estimation in case-referent studies. Am J Epidemiol 103:226–235

Miettinen OS, Cook EF (1981) Confounding: essend and detection. Am J Epidemiol 114:593–603

Morgenstern H (1982) Uses of ecologic analysis in epidemiological research. Am J Public Health 72:1336–1344

Morgenstern H, Bursic ES (1982) A method for using epidemiologic data to estimate the potential impact of an intervention on the health status of a target population. J Community Health 7:292–309

Morgenstern H, Kleinbaum D, Kupper LL (1980) Measures of disease incidence used in epidemiologic research. Int J Epidemiol 9:97–104

Neuhaus JM, Jewell NP (1990) The effect of retrospective sampling on binary regression models for clustered data. Biometrics 46:977–990

Neuhaus JM, Kalbfleisch JD, Hauck WW (1991) A comparison of cluster-specific and population-averaged approaches for correlated binary data. Int Stat Rev 59:25–35

Neutra RR, Drolette ME (1978) Estimating exposure-specific disease rates from case-control studies using Bayes' theorem. Am J Epidemiol 108:214–222

Oakes D (1981) Survival times: aspects of partial likelihood (with discussion). Int Stat Rev 49:235–264

Ouellet BL, Romeder JM, Lance JM (1979) Premature mortality attributable to smoking and hazardous drinking in Canada. Am J Epidemiol 109:451–463

Palta M (2003) Quantitative methods in population health: extensions of ordinary regression. Wiley, Hoboken

Palta M, Lin C-Y (1999) Latent variables, measurement error and methods for analyzing longitudinal binary and ordinal data. Stat Med 18:385–396

Palta M, Lin C-Y, Chao W (1997) Effect of confounding and other misspecification in models for longitudinal data. In: Modeling longitudinal and spatially correlated data. Lecture notes in statistics series 122. Proceeding of the Nantucket conference on longitudinal and correlated data. Springer, Heidelberg/New York, pp 77–88

Prentice RL, Breslow NE (1978) Retrospective studies and failure time models. Biometrika 65:153–158

Prentice RL, Pyke R (1979) Logistic disease incidence models and case-control studies. Biometrika 66:403–411

Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE (1978) The analysis of failure times in the presence of competing risks. Biometrics 34:541–554

Rao CR (1965) Linear statistical inference and its application. Wiley, New York, pp 319–322

Ries LAG, Eisner MP, Kosary CL, Hankey BF, Miller BA, Clegg L, Mariotto A, Feuer EJ, Edwards BK (eds) (2004) SEER cancer statistics review, 1975–2001. National Cancer Institute. Bethesda. http://seer.cancer.gov/csr/1975_2001. Accessed 21 May 2004

Robins JM, Greenland S (1989) Estimability and estimation of excess and etiologic fractions. Stat Med 8:845–859

Rockhill B, Newman B, Weinberg C (1998a) Use and misuse of population attributable fractions. Am J Public Health 88:15–21

Rockhill B, Weinberg C, Newman B (1998b) Population attributable fraction estimation for established breast cancer risk factors: considering the issues of high prevalence and unmodifyability. Am J Epidemiol 147:826–833

Rothman KJ, Greenland S (1998) Modern epidemiology. Lippincott-Raven, Philadelphia

SAS Institute Inc (1999) SAS/STAT user's guide. Version 8. SAS Institute Inc, Cary

Schlesselman JJ (1982) Case-control studies. Design, conduct and analysis. Oxford University Press, New York

Seiler FA, Scott BR (1987) Mixtures of toxic agents and attributable risk calculations. Risk Analysis 7:81–90

Seiler FA, Scott BR (1986) Attributable risk, probability of causation, assigned shares, and uncertainty. Environ Int 12:635–641

Siemiatycki J, Wacholder S, Dewar R, Cardis E, Greenwood C, Richardson L (1988) Degree of confounding bias related to smoking, ethnic group and SES in estimates of the associations between occupation and cancer. J Occup Med 30:617–625

Smith L (1998) Person-years of life lost. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics. Wiley, Chichester, pp 3324–3325

Spiegelman D, Colditz GA, Hunter D, Hetrzmark E (1994) Validation of the Gail et al model for predicting individual breast cancer risk. J Natl Cancer Inst 86:600–607

Sturmans F, Mulder PGH, Walkenburg HA (1977) Estimation of the possible effect of interventive measures in the area of ischemic heart diseases by the attributable risk percentage. Am J Epidemiol 105:281–289

Tarone RE (1981) On summary estimators of relative risk. J Chronic Dis 34:463–468

Tsiatis AA (1981) A large-sample study of Cox's regression model. Ann Stat 9:93–108

Tuyns AJ, Pequignot G, Jensen OM (1977) Le cancer de l'œsophage en Ille-et-Vilaine en fonction des niveaux de consommation d'alcool et de tabac. Bull Cancer 64:45–60

US National Cancer Institute (2004) Breast cancer risk assessment tool. An interactive tool to measure a woman's risk of invasive breast cancer. http://bcra.nci.nih.gov/brc. Accessed 12 May 2004

Uter W, Pfahlberg A (1999) The concept of attributable risk in epidemiological practice. Biom J 41:985–999

Uter W, Pfahlberg A (2001) The application of methods to quantify attributable risk in medical practice. Stat Methods Med Res 10:231–237

Wacholder S, Benichou J, Heineman EF, Hartge P, Hoover RN (1994) Attributable risk: advantages of a broad definition of exposure. Am J Epidemiol 140:303–309

Wahrendorf J (1987) An estimate of the proportion of colo-rectal and stomach cancers which might be prevented by certain changes in dietary habits. Int J Cancer 40:625–628

Walter SD (1975) The distribution of Levin's measure of attributable risk. Biometrika 62:371–374

Walter SD (1976) The estimation and interpretation of attributable risk in health research. Biometrics 32:829–849

Walter SD (1980) Prevention for multifactorial diseases. Am J Epidemiol 112:409–416

Walter SD (1983) Effects of interaction, confounding and observational error on attributable risk estimation. Am J Epidemiol 117:598–604

Whittemore AS (1982) Statistical methods for estimating attributable risk from retrospective data. Stat Med 1:229–243

Whittemore AS (1983) Estimating attributable risk from case-control studies. Am J Epidemiol 117:76–85

Wooldridge JM (2001) Econometric analysis of cross section and panel data. MIT, Boston

Woolf B (1955) On estimating the relationship between blood group and disease. Ann Hum Genet 19:251–253

Wu K, Brown P (2003) Is low-dose Tamoxifen useful for the treatment and prevention of breast cancer (editorial)? J Natl Cancer Inst 95:766–767

Zhang J, Yu KF (1998) What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. JAMA 280:1690–1691

Descriptive Studies 5

Freddie Bray and D. Maxwell Parkin

Contents

F. Bray (✉)
Section of Cancer Information, International Agency for Research on Cancer, Lyon, France

D.M. Parkin
Clinical Trials Service Unit & Epidemiological Studies Unit, Department of Medicine, University of Oxford, Oxford, UK

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_4,

5.1 Introduction

We begin by setting out some definitions in descriptive epidemiology, the sources of data from which such studies arise and provide a brief outline of the sections that comprise this chapter.

5.1.1 Definitions

A distinction has traditionally been drawn between “descriptive” and “analytical” epidemiology, and their characteristics as “hypothesis generating” or “hypothesis testing,” respectively, have been taught to generations of students. This distinction may perhaps reflect the part separation of the distribution from the causes of disease, as in the current dictionary definition of epidemiology (Porta 2008):

> The study of the occurrence and distribution of health-related states or events in specified populations, including the study of the determinants influencing such states, and the application of this knowledge to control the health problems.

Describing the distribution of disease is an integral part of the planning and evaluation of health-care services, but, in the context of investigative epidemiology, the distinction is an arbitrary one. There are no real differences in the concepts, methods, or deductive processes between descriptive and analytical epidemiology, for example, between the information conveyed by the observations of an association between the risk of liver cancer and being engaged in a specific occupation, or having markers of infection by a certain virus. Both may tell us something about the cause of liver cancer. Only the sources of information differ. In the former case, it has derived from some routine source (a dataset or register maintained for general disease surveillance purposes or even for unrelated administrative reasons). These sources include information on possible exposures or disease outcomes that have not been collected with the testing of any specific hypothesis in mind. It is this use of routinely collected data that characterizes descriptive studies.

Descriptive epidemiology is certainly not synonymous with ecological studies of groups, as suggested by some authors (Estève et al. 1994), since most “descriptive” studies have information on distribution and levels of several exposure variables for individual members of the population studied. In their classic textbook, MacMahon and Trichopoulos (1996) liken this phase of epidemiological investigation to the early questions in the parlor game “20 questions” – using generally available information to focus down the field of inquiry to one that may need expensive ad hoc study. But the variables available in routine data sources are no less “exposures” from the point of view of methodology, or deductive reasoning, than are those measured by questionnaire, physical examination, or biochemical tests. The fact that some “exposures” may be remote from the molecular mechanisms involved in disease etiology is a familiar one when considering “cause” of disease, especially from a public health perspective of devising appropriate methods of prevention. “Cause” is a relative concept that only has meaning in epidemiology terms of its removal being associated with a diminished risk of the disease, and, in this context, it is just as relevant to improve educational levels in a population as a means of reducing infection by HIV as it is to identify the mechanisms by which the virus enters the host cell.

"Exposures" in descriptive epidemiology are those characteristics of individuals that are present in the preexisting datasets available for study. The most commonly available are personal characteristics, the so-called "demographic" variables, systematically collected by vital statistics and health-care institutions. They include birth date (or age), sex, address (current place of residence), birthplace, race/ethnic group, marital status, religion, occupation, and education. From sources within the health sector, there may be much more detail on diagnostic and therapeutic interventions, while community surveys may include information on health determinants, such as tobacco and alcohol use, weight, height, and blood pressure.

Disease outcomes may be in terms of incident cases (from disease registers), deaths (vital statistics), episodes of morbidity (utilization statistics from health services), or prevalence of, especially, chronic conditions from population surveys.

Information on "exposure" and "outcome" for the same individual may be taken from a single source (e.g., a disease register or population survey), or it may be necessary to perform record linkage between different sources to obtain exposure and outcome information on individuals in the population under study.

5.1.2 Sources of Data

There is a wide variety of sources of information that can be drawn upon for information on exposure and disease outcome. They are of two broad types: systems based on populations, containing data collected through personal interviews or examinations, and systems based on records, containing data collected from vital and medical records. They include:

- Census data or population registers
- Medical birth registries (health events related to pregnancy and birth)
- Vital statistics (especially death certificate data)
- Disease registers, recording new cases of specific diseases in defined populations (such as registries that collect information on diagnoses of cancer, insulin-managed diabetes mellitus, etc.)
- Notification systems (especially for infectious diseases)
- Hospital activity statistics, especially on admissions/discharges from hospital, including diagnosis
- Primary care contacts
- Diagnostic services (pathology)
- Community surveys (e.g., those carried out by the National Center for Health Statistics in the USA (Curtin et al. 2012))

Some of these are described in chapters ►Use of Health Registers and ►Emergency and Disaster Health Surveillance of this handbook.

5.1.3 Outline of the Chapter

The chapter comprises four parts, beginning with an introduction of the most important measurements in descriptive epidemiology in Sect. 5.2. Since these are primarily concerned with risk or burden of disease in a single population, appropriate methods for comparisons between populations – the hallmark of epidemiology – are presented in Sect. 5.3. Section 5.4 illustrates how these tools can be applied in the study designs familiar to epidemiologists, with a special emphasis on ecological studies, while Sect. 5.5 provides a series of examples, illustrating the principles of descriptive studies.

5.2 Measurement

5.2.1 Incidence

Incidence is the number of new cases occurring. It can be expressed as an absolute number of cases or in relation to the size of the population at risk and time during which the cases occur, as an incidence rate (cf. chapter ▸Rates, Risks, Measures of Association and Impact of this handbook). Incidence requires definition of the moment at which the disease "begins," when an individual becomes a new "case." While this may be relatively straightforward for many infectious diseases, for most acute and chronic diseases, the time of onset is less clear cut and is by convention considered to be the time of diagnosis. This is necessarily a somewhat arbitrary point in time and dependent upon local circumstances. Incidence may refer to the number of new disease events or to the number of individuals affected. The distinction is important where the same individual may have more than one event of the same disease, during the period of observation (e.g., common infections and accidents).

Incidence data are available from disease registers as well as from notification systems for infectious diseases. Disease registers are part of a surveillance system for various diseases, and an accurate registration from a notification of disease events depends upon the patient seeking medical advice, the correct diagnosis being made, and notification of it being made to the public health authorities. Completeness is very variable but probably higher for serious or highly contagious diseases. Registers have been more important, and successful, for cancer than for any other condition. This is because of the serious nature of most cancers, which means that, except in certain low-resource countries with limited access to medical care and concepts, the vast majority of those affected will present for diagnosis (and treatment, if available). As a result, enumeration of incident cases of cancer is relatively easy in comparison with other diseases, and this has permitted the establishment of a worldwide network of cancer registries, providing data on defined populations (Parkin 2006). Incidence data from cancer registries worldwide that meet defined criteria regarding completeness and validity are published at 5-year intervals in the *Cancer Incidence in Five Continents* series. The latest volume (9th) contains comparable incidence information from 225 registries in 60 countries, mainly over the period 1998–2002 (Curado et al. 2007).

Other sources of information on incidence include:

(a) Retrospective or prospective surveys for incident cases of a particular disease. The approach is similar to ongoing registers, except that it is limited in time.
(b) Community surveys. General morbidity surveys record all cases of disease appearing in a sample of the community, for example, at primary care level, during a period of time. They are most efficient for relatively common conditions.
(c) Hospital activity statistics. These summarize hospital admissions. Such statistics are usually *event*-based, so there may be multiple admissions for the same disease event. The number of hospitalizations is also affected by the facilities available, admissions policies, and social factors. It may be difficult to define the population at risk, unless national-level data can be compiled.

Incidence rates are of particular value in the study of disease etiology, since they are informative about the risk of developing the disease in different population groups.

5.2.2 Prevalence

Prevalence is the number or, more usually, the proportion of a population that has the disease at a given point in time (Rothman and Greenland 1998; cf. chapter ▸Rates, Risks, Measures of Association and Impact of this handbook). For many diseases (e.g., hypertension, diabetes), prevalence usefully describes the number of individuals requiring care and may be useful in planning health services. Prevalence is proportional to the incidence of the disease and its duration (and when both are constant, then **P**revalence = **I**ncidence × **D**uration). In the absence of useful data on incidence or mortality, prevalence may be used to compare the risk of disease between populations, although this has clear drawbacks, not least because prevalence is related also to the mean duration of the disease. The mean duration may simply reflect the availability of screening and treatment strategies, allowing prolonged survival. For some chronic diseases, what to consider as the moment of cure (after which an individual is no longer a "case") is a problem when trying to calculate comparable indices of prevalence between populations (Bray et al. 2013). Population surveys are a common source of information on prevalence of the more common conditions or complaints, including those that may exist in asymptomatic form, and therefore remain unrecognized in a clinical setting.

A less commonly used measure is *period prevalence*, which is the sum of all cases of the disease that have existed during a given period divided by the average population at risk during the period. It has been used for studying mental illness, where the exact time of onset is difficult to define and when it may be difficult to know whether the condition was present at a particular point in time.

5.2.3 Case Fatality/Survival

Case fatality is a measure of the severity of a disease. It is the proportion of cases of a particular disease that are fatal within a specified time. *Survival* is proportion of cases that do not die in a given interval after diagnosis (and is equal to *1-fatality*). The survival time is defined as the time that elapsed between diagnosis and death. Computation of survival depends upon follow-up of diagnosed patients for deaths or withdrawal from observation. There are two related approaches to the estimation of survival: the Kaplan-Meier and actuarial, or life-table, methods. The former (Kaplan and Meier 1958) is particularly useful when exact survival times are available, since smooth estimates of survival as a function of time since diagnosis can be obtained. The actuarial method requires a life-table with survival times grouped usually into intervals that permit the calculation of the cumulative probability of survival at time t_i from the conditional probabilities of survival during consecutive intervals of follow-up time up to and including t_i (Cutler and Ederer 1958; Ederer et al. 1961). Information from all cases is used in the estimation of survival, including those withdrawn due their follow-up ending owing to closure of study and those who are lost to follow-up before the termination. In both cases, follow-up is censored before the time of the outcome event, usually the death of the patient. "Observed survival" is influenced not only by mortality from the disease of interest but also by deaths from other causes. If these deaths can be identified, they can be treated as withdrawals, and the "corrected survival" (also referred to as "net survival") calculated. Alternatively, allowance for deaths due to causes other than the disease under study is made by calculation of "relative survival" (Ederer et al. 1961). This makes use of an appropriate population life-table to estimate expected numbers of deaths. The issue of competing risks is discussed in detail in chapter ▸Rates, Risks, Measures of Association and Impact of this handbook. For comparisons between different populations, age standardization of survival is necessary (see Sect. 5.3.1.1).

For a more detailed description of survival analysis, see chapter ▸Survival Analysis of this handbook.

5.2.4 Mortality

Mortality is the number of deaths occurring in a population and is the product of the incidence and the fatality of the disease.

Mortality statistics derive from the information on death certificates, collected by civil registration systems recording vital events (births, marriages, deaths). The responsible authority varies between countries, but usually the first level of data collection and processing is the municipality or province, with collation of national statistics being the responsibility of the Ministry of Health or the Interior Ministry. Death certificates record information on the person dying, and the cause of death, as certified, usually by a medical practitioner. The International Classification of Diseases (ICD) provides a uniform system of nomenclature and coding and a recommended format for the death certificate (WHO (World Health Organization) 1992). Mortality statistics are produced according to the underlying cause of death; this may not equate with the presence of a particular disease. Although the ICD contains a set of rules and guidelines that allow the underlying cause to be selected in a uniform manner, interpretation of the concept probably varies considerably, for example, when death occurs from pneumonia in a person previously diagnosed as having cancer. Comprehensive mortality statistics thus require that diagnostic data are available on decedents, which are transferred in a logical, standardized fashion to death certificates, which are then accurately and consistently coded, compiled, and analyzed.

It is well known that mortality data may be deficient both in completeness of ascertainment and in the validity of the recorded cause of death (Alderson 1981). A huge number of studies of the validity of cause-of-death statements in vital statistics data have been carried out. These compare cause of death entered on the death certificate with a reference diagnosis, which may be derived from autopsy reports (e.g., Heasman and Lipworth 1966), detailed clinical records (Puffer and Wynne-Griffith 1967), or cancer registry data (Percy et al. 1981). They reveal that the degree of accuracy of the stated cause of death declines as the degree of precision in the diagnosis increases. Despite the problems, mortality data remain an extremely valuable source of information on disease burden and a useful proxy for risk of disease in many circumstances. A major advantage is in their widespread availability. According to *World Health Statistics 2012* (WHO 2012), only 34 countries (representing 15% of the global population) produce high-quality cause-of-death data (predominantly in Europe and the Americas), while countries representing almost two-thirds of the world population produce lower-quality cause-of-death data, with the remaining 74 countries – mainly low-resource countries – lack such data altogether. National-level statistics are collated and made available by the WHO (2012); however, the fact of publication by national and international authorities is not a guarantee of data quality. For some countries, or time periods, coverage of the population is manifestly incomplete, and the so-called mortality rates produced are implausibly low. The WHO Statistical Information System (WHOSIS) (now incorporated into the Global Health Observatory (GHO)) publishes tables of estimated coverage and completeness of mortality statistics in their database (Mathers et al. 2005). As well as deficiencies in these, quality of cause-of-death information may be poor. This can sometimes be predicted when a substantial proportion of certificates are completed by non-medical practitioners. (WHO formerly published a useful table in "World Health Statistics Annual" (WHO 1998) giving – for a few countries at least – the relevant percentage.) Otherwise, quality of data must be judged from indicators such as the proportion of deaths coded to *Senility and Ill Defined Conditions*.

The mortality rate – the number of deaths in relation to the population at risk, in a specified period of time – provides an indicator of the average risk to the *population* of dying from a given cause (while fatality represents the probability that an *individual* with the disease will die from it). Mortality rates are the most useful measure of the impact or burden of disease in a population. They are probably equally often used as a convenient proxy measure of the risk of acquiring the disease (incidence) when comparing different groups, because of their availability, although such use introduces the assumption of equal survival/fatality in the populations being compared.

The *infant mortality rate* is a widely used indicator of the level of health and development. It is calculated as the number of deaths in children aged less than one, during a given year, divided by the number of live births in the same year. Other similar indicators include the fetal death rate, stillbirth or late fetal death rate, perinatal mortality rate, neonatal mortality rate, and postneonatal mortality rate. These data may be collected by household survey, rather than the vital statistics system, especially in developing countries (United Nations 1984).

5.2.5 Person-Years of Life Lost

The concept of person-years of life lost (PYLL) was introduced over 50 years ago (Dempsey 1947) in order to refine the traditional mortality rates, by providing a weighting for deaths occurring at different ages. These methods started to become more widely used from the late 1970s in health services planning (Murray and Axtell 1974). There are many variations in the calculations used (summarized by Murray 1994). There are four variants of "normal" lifespan against which to compare premature death. The simplest is to choose a fixed value for the potential limit where ages ranging from 60 to 85 have been used. *Potential years of life lost* are calculated as

$$\sum_{x=0}^{L} d_x (L - x)$$

where d_x is the number of deaths at age x and L is the potential limit to life. This method gives no importance to deaths over the upper limit. To avoid this, calculating the *expected years of life lost*, using the expectation of life at the age of death (e_x), derived from an appropriate life-table, seems a better solution:

$$\sum_{x=0}^{l} d_x e_x$$

where l is the last age group and e_x is the expectation of life at each age. The expectation of life (e_x) may be taken from a period life-table or, more appropriately, be cohort-specific. This method is not suitable, however, for comparing between populations with different expectations of life. For this purpose, a standard expectation of life (e_x^*), taken from some ideal standard population, should be used. The calculated indicator then becomes the *standard expected years of life lost*. Another variation is to give different weights to years of life lost at different ages. The rationale is that the economic, or social, "value" of individuals varies with their age.

In addition, discounting may be used to give decreasing weights to the life years saved over time, admitting that life years in the future are valued less highly than at present (Das Gupta et al. 1972; Layard and Glaister 1994).

The approach has been taken a step further, with the development of indices that take into account non-fatal health outcomes of disease, such as *quality-adjusted life years (lost)* (*QALYs*) or *disability-adjusted life years (lost)* (*DALYs*) (Murray 1994; Morrow and Bryant 1995). Essentially, these admit that, between onset of a disease and death or recovery, there is a spectrum of morbidity, which can be quantified in terms of its duration and severity. Three elements are needed in calculating these indices, therefore: the incidence of the disease, its mean duration or, equivalently, survival probability, and a measure of life "quality" in between onset and end of disease. The problem in using these indices lies in ascribing values to quality of life, or level of disability, since both are subjective and will vary with time since diagnosis and in different cultural and socioeconomic settings. Nevertheless, the estimation of DALYs for different conditions worldwide has been widely used by the World Health Organization (WHO) as a means of establishing priorities for health-care programs (WHO 2000). The calculation involves summing the standard expected years of life lost (using a standard model life-table with expectation at life at birth being 82.5 for females and 80 for males), with the time lived between onset and death in different disability classes having a severity weighting between 0 and 1. Some examples of long-term disease and injury sequelae falling into each class are given in Table 5.1; the full list can be found in WHO (2008). Both an age weighting function and a discount rate (of 3% per year) are applied (Murray and Lopez 1996). The International Agency for Research on Cancer (IARC) has recently proposed a framework (Soerjomataram et al. 2012a) to estimate DALYs due to cancer in 184 countries (Soerjomataram et al. 2012b).

5.2.6 Rates of Disease

Although the dimensions of a health problem may be expressed by the absolute numbers of events e.g., the numbers of cases of infectious diseases, including AIDS, reported through WHO's surveillance data (WHO 2012), comparisons between population groups require that the number of events is related to the size of the population in which they occur, by the calculation of rates or proportions. "Rate," as the name implies, incorporates a time dimension – the number of events occurring in a defined population during a defined time period. The term rate is often used interchangeably with *risk*, although the risk of disease is a probability, or proportion, and describes the accumulation of the effect of rates over a given period of time. Ideally, we would estimate a rate by ascertaining, for every individual in the population, the risk of being diagnosed or dying at a given age and specific point in time. This *instantaneous rate* requires that the designated period of time is infinitely small, approaching zero. In practice, the average rate of occurrence of new cases or deaths in a sufficiently large population is calculated for a sufficiently long time period. In this formulation, the denominator is the underlying *person-time at risk* in

Table 5.1 Disability classes for the Global Burden of Disease study, with examples of long-term disease and injury sequelae falling into each class (WHO 2008)

Disability class	Severity weights	Conditions
I	0.00–0.02	Stunting due to malnutrition, schistosomiasis infection, long-term scarring due to burns (less than 20% of body)
II	0.02–0.12	Amputated finger, asthma case, edentulism, mastectomy, severe anemia, stress incontinence
III	0.12–0.24	Angina, HIV not progressed to AIDS, infertility, alcohol dependence and problem use, low vision (<6/18, >3/60), rheumatoid arthritis
IV	0.24–0.36	Amputated arm, congestive heart failure, deafness, drug dependence, Parkinson disease, tuberculosis
V	0.36–0.50	Bipolar affective disorder, mild mental retardation, neurological sequelae of malaria, recto-vaginal fistula
VI	0.50–0.70	AIDS cases not on antiretroviral drugs, Alzheimer and other dementias, blindness, Down syndrome
VII	0.70–1.00	Active psychosis, severe depression, severe migraine, quadriplegia, terminal stage cancer

which the cases or deaths in the numerator arose. Prevalence and survival, as defined above, are proportions (or percentages), not rates, as there is no time dimension, although the term "rate" is frequently appended.

5.2.7 Population at Risk

For the denominators of proportions, such as prevalence and survival, the population at risk comprises the number of individuals at a point in time. When rates are calculated, the time period of observation also needs to be specified. The denominator is calculated as units of person-time (Last 2001), whereby each person in the study population contributes one person-year for each year of observation before disease develops or the person is lost to follow-up. In prospective cohort studies, the individuals can be followed up until the end of the study, and summation of the varying lengths of individual follow-up accurately represents the person-time at risk of disease. However, in descriptive epidemiology, information on the population-at-risk is not usually available at the individual level, unless there is an accurate population register. It is usually necessary to approximate person-years at risk using cross-sectional population data from national statistical organizations. The estimation of the denominator, a summation of the midyear estimates for each of the years under consideration, is thus dependent on both the availability and completeness of demographic information on the population under study.

In most developing and developed countries, 10-yearly population censuses provide basic population estimates by age, sex, and census year, and statistical offices often produce estimates for intercensal years, based on rates of birth, death,

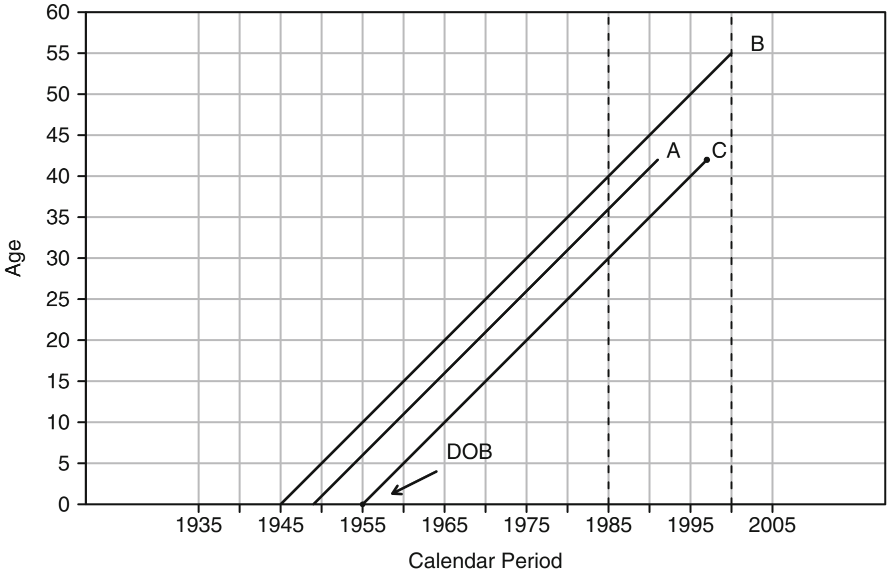


Fig. 5.1 The modern Lexis diagram depicting three subjects under observation between 1985 and 2000. Subject *A* is lost to follow-up aged 42 in 1991, *B* is alive at the end of follow-up (end of 2000), and *C* has the outcome of interest at the age of 42 in 1997 – the date of birth (*DOB*) can be ascertained for *C* by period – age = 1955, and his/her "life journey" represented by the diagonal line from *DOB* to 1997

and migration. The approximation assumes there is stability in the underlying population, as individuals traverse the age-time plane represented by the well-known Lexis diagram (Fig. 5.1). The German demographer Lexis (1875) described this graphical representation of the life history of subjects according to birth cohort (the abscissa) and age (the ordinate). The modern interpretation of the Lexis diagram displays subjects arranged by calendar year of event and age at time of event on the same unit scale, with each cell corresponding to a year of birth, the diagonal tracing the experience of subjects born in the same year, who are under observation until either the end of follow-up; the event of interest occurs; or they are lost to follow-up. In the commonly tabulated system used for vital rates, a synthetic birth cohort over a 10-year range is derived from the combined experience of a 5-year age group over 5 years of occurrence. Approximate cohorts are identified by their central year of birth and overlap each other by exactly 5 years. Given a steady state of demographic gains and losses, where the number of individuals in a designated period entering an age group equals the number that leave, the method can be considered to provide adequate estimates of the person-time at risk in most circumstances. For diseases that are relatively rare (of low prevalence), the estimate is not unduly biased by the fact that the numerator is a subset of the denominator.

It should be recalled that the number of units of person-time does not represent a number of independent observations. One hundred person-years of observation may result from 100 persons being followed for 1 year each or from 20 people under observation for 5 years.

Population at risk should, ideally, only include those persons who are potentially susceptible to the disease being studied. Sometimes, this is taken into account in the denominator, for example, in studying occupational diseases (where only those in the relevant occupation are at risk), and in infectious disease epidemiology, where a large proportion of the population is immune.

5.3 Comparisons Between Populations

In descriptive studies, comparisons between two or more populations are usually based on rates that account for the person-time at risk in each group using mid-year population estimates. The summary measure is estimated by stratifying on some well-known factor to remove its effect as a potential confounder. Comparisons of disease risk as a rule must take into account the effects of age, given its influence on both the disease process and the exposure of interest across groups.

The graphical representation of age-specific rates in a particular population describes how risk evolves with age. In comparing age-disease curves in two (or more) populations, it is a useful point of departure, often alerting the researcher to anomalies in the data or to certain hypotheses that warrant further investigation. Standard methods exist for comparing rates in two or more populations. The simple techniques include the comparison of rates based on age standardization and stratification methods that pool the age-specific rates to obtain a weighted ratio (Mantel and Haenszel 1959). In the presence of heterogeneity in the age-specific rates, pooling or standardization may not provide a satisfactory measure of relative risk, and visual inspections together with statistical tests that look for departures from the assumption of proportionality (homogeneity) should be investigated (Estève et al. 1994).

Greater flexibility and a more unified framework is a prerequisite when dealing with a series of population comparisons, and, given most diseases are multifactorial, the analysis must consider the association between numerous factors on disease risk while adjusting for several potential confounders. Statistical models offer quantitative and comparable estimates of disease risk based on objective criteria for choosing the best description of the data and whether observed variations are real or due to chance. Some methods for the analysis of two groups and multiple groups are briefly described below.

5.3.1 Comparisons Between Two Groups

5.3.1.1 Standardization

The Direct Method Crude rates can be thought of as a weighted sum of age-specific rates that render *biased* comparisons between populations if the weights that represent the size of each age stratum are different in each population compared. The age-standardized rate is the summary rate that would have been observed, given the schedule of age-specific rates, in a population with the age composition of some reference population, called the *standard*. The calculation of the standardized rate is an example of *direct standardization*, whereby the observed age-specific rates in each group are applied to the *same* standard, that is, the same age-specific weights. Age groups are indexed by the subscript i, d_i is the number of cases, y_i is the number of person-time at risk (frequently obtained by multiplying population estimates based on those at risk by the length of the observation period), and w_i is the proportion of persons or *weight* of age group i in the chosen standard population. The age-standardized rate (*ASR*) is given by

$$ASR = \sum_i d_i w_i / y_i .$$

The main criticism of the technique stems from the need to select an arbitrary standard population. The most widely used reference for global comparisons is the world standard, as proposed by Segi (1960) on the basis of the pooled population of 46 countries and modified for the first volume of *Cancer Incidence in Five Continents* by Doll et al. (1966). Although this does not really resemble the age structure of the current population of the world (so that *ASR*s will rarely be similar to crude rates), this is of little importance, since it is the *ratio* of *ASR*s (the standardized rate ratio), an estimate of relative risk between populations, that is the focus of interest. This has been shown to be quite insensitive to the choice of standard (Bray et al. 2002; Gillum 2002).

Another form of direct standardization involves the cumulative risk, defined as the probability that an individual will develop the disease in question during a certain age span, in the absence of other competing causes of death. The age span over which the risk is accumulated must be specified. The age ranges 0–64 and 0–74 are generally used and attempt to give two representations of the lifetime risk of developing the disease. Other age ranges may be more appropriate for more specific needs, such as investigating childhood diseases. If the cumulative risk using the above age ranges is less than 10%, as is the case for relatively rare diseases, it can be approximated very well by the cumulative rate (cf. chapter ▸Rates, Risks, Measures of Association and Impact of this handbook).

The cumulative rate is the summation of the age-specific rates over each year of age from birth to a defined upper age limit. As age-specific incidence rates are typically computed for 5-year age intervals, the cumulative rate is then five times the sum of the age-specific rates calculated over the 5-year age groups, assuming the age-specific rates are the same for all ages within the 5-year age stratum. The cumulative rate from 0 to 74 is given by

$$5 \sum_{i=1}^{15} d_i / y_i .$$

The precise mathematical relationship between the cumulative rate and the cumulative risk (Day 1992) is

$$\text{cumulative risk} = 1 - \exp(-\text{cumulative rate}).$$

The cumulative rate has several advantages over age-standardized rates. Firstly, the predicament of choosing an arbitrary reference population is irrelevant. Secondly, as an approximation to the cumulative risk, it has a greater intuitive appeal and is more directly interpretable as a measurement of lifetime risk, assuming no other causes of death are in operation.

The Indirect Method An alternative form of age standardization, known as the *indirect method*, involves calculating the ratio of the total number of cases *observed* in the population of interest, $O = \sum_i d_i$, to the number of cases which would be *expected*, E, if the age-specific risks of some reference population applied. The expected number of cases in a study population is given by

$$E = \sum_i m_i y_i$$

where the m_i are age-specific rates for the reference population and y_i the number of persons in age class i in the population of interest. The ratio is termed the standardized mortality ratio (SMR) when deaths and mortality rates are used; the terms "standardized incidence ratio" (SIR) or "standardized morbidity ratio" are used for incidence data. Expressed as a percentage, the calculation is

$$SMR \text{ (or } SIR) = \frac{O}{E} \times 100.$$

There are two ways in which the reference population can be chosen. If the aim is to compare several populations with a specific reference population, then it would be sensible to choose a reference population with relatively large numbers of observed cases, since this increases the precision of the reference rates. A second strategy would be to create a pooled population from those to be compared; this has the advantage of increasing the precision of the reference rates and is analogous to comparing the observed rate in each population with that expected if the true age-specific risks were identical in all of the populations. Whichever approach is taken, it is important to realize that the *SMR*s of the individual populations can only be compared with the reference population. In addition, the *SMR* for population A compared with population B is *not* the inverse of the *SMR* for B compared with A.

Both direct and indirect standardization can give a reasonable summary of a multiplicative effect and are normally close in practice. Indirect standardization however requires a further assumption of uniformity of the effect – the *SMR* has optimal statistical properties only if the force of incidence in the population of interest is *proportional* to that of the reference.

Direct standardization is preferred for statistical reasons, given that a ratio of *SMR*s for two comparison groups may in some instances misrepresent the underlying stratum-specific ratios (Breslow and Day 1987). The risk ratio of two directly standardized rates (based on the same standard) has an associated confidence interval based on either exact variance calculations (Rothman and Greenland 1998) or a close approximation given by Smith (1987). As already mentioned, heterogeneity in the age-specific rates may render the relative risk estimate invalid.

5.3.1.2 Stratification: The Mantel-Haenszel (MH) Estimate

The standardized rate ratio is not a very efficient estimator of relative risk, since the weightings for the age strata are entirely arbitrary. More efficient summary measures assume uniformity across strata. The common rate ratio ρ for population 1 compared with population 0 is $\rho = \lambda_{1i}/\lambda_{0i}$. This relation can be written as $\lambda_{1i} = \lambda_{0i} \cdot \rho$ or, in the form of a *log-linear model*, as

$$\ln(\lambda_{1i}) = \ln(\lambda_{0i}) + \ln(\rho)\,.$$

This is sometimes called the *multiplicative model*.

The Mantel-Haenszel (MH) estimate of the common rate ratio ρ is simply the weighted average of the ratio of the age-specific rates:

$$\hat{\rho} = \frac{\sum\limits_i \frac{d_{1i}\, y_{0i}}{y_{\cdot i}}}{\sum\limits_i \frac{d_{0i}\, y_{1i}}{y_{\cdot i}}}$$

where $\cdot$ denotes summation over the index it replaces.

The Cochran-Mantel-Haenszel (CMH) test tests the hypothesis $\rho \neq 1$ against the null hypothesis $\rho = 1$ (Cochran 1954), for example, whether the force of incidence is identical in the two populations being compared ($\rho = 1$). However, the CMH test is valid only if the age-specific rate ratios are approximately proportional, that is, $\rho_i \approx \rho$. Therefore, it is important to check this assumption. Several such tests are described by Breslow (1984); one method involves comparing the numbers of cases observed and expected under the assumption of proportionality, while another tests the same null hypothesis against the specific alternative of a trend (increasing or decreasing) in the age-specific rate ratios. The different hypotheses which can be tested and their corresponding alternative hypotheses are shown in Table 5.2 and Fig. 5.2 below. More details and the required formulae can be found in Estève et al. (1994).

Table 5.2 Null hypotheses and corresponding alternatives for the common rate ratio ρ

H_0 (null hypothesis)	H_a (alternative hypothesis)
$\rho = 1$	$\rho \neq 1$, test for a common rate ratio (i.e., assumes proportionality)
$\rho_i = \rho$	ρ_i unrestricted, test for heterogeneity in the ρ_i
$\rho_i = \rho$	$\rho_i = \rho \times f(i)$, test for trend in the ρ_i with the age values i

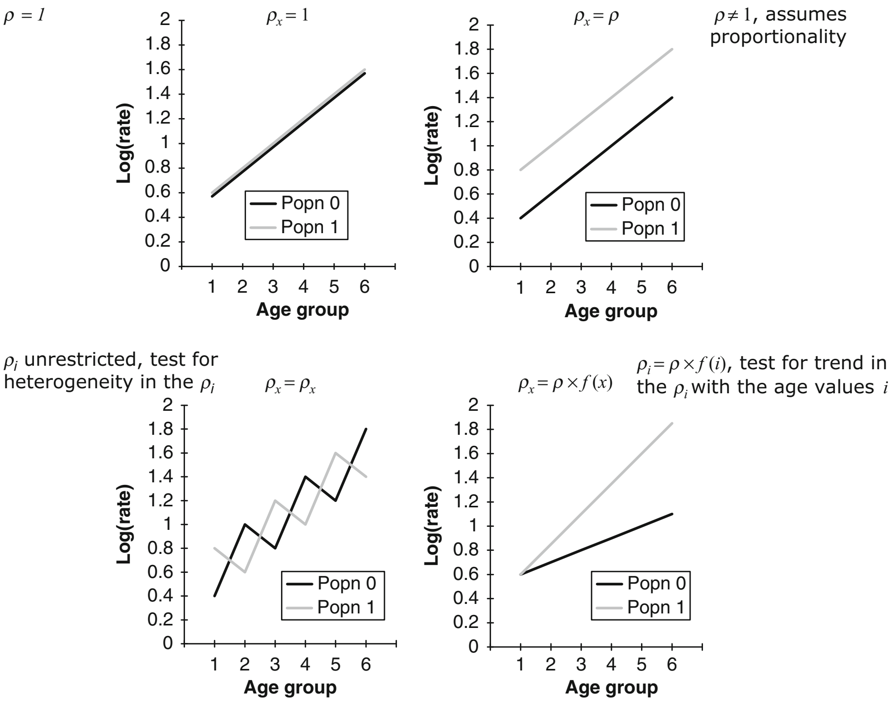


Fig. 5.2 Graphical representation of the null hypotheses and alternatives described in Table 5.2

5.3.2 Comparisons Between Multiple Groups

In practice, when comparing multiple populations, the simple methods above offer some serious limitations. Generally, a series of pairwise comparisons may yield spurious significant results due to multiple testing and is therefore not appropriate. In indirect standardization, the choosing of the age-specific risks of one population as a reference over several others is inconsistent: the *ratio* of population 1 relative to population 0 is not the inverse of the *ratio* of the population 0 relative to population 1.

In most studies, there are a number of confounders that require examination and possible adjustment, other than age. The Mantel-Haenszel estimates may be extended to adjust simultaneously for several confounders, but we may not have sufficient data to simultaneously consider many strata. In addition, it is not possible to classify an explanatory variable as an exposure or a confounder using this method, and a separate analysis is required to obtain rate ratios for each exposure adjusted for the confounders.

A regression model provides a uniform framework for estimating the magnitude of the effect of interest, testing whether the effect is uniform across subgroups of the populations (effect modification), whether the effects of potential confounders may account for the effect, and whether a particular model is given a parsimonious but adequate description of the observed data.

Further details on appropriate regression models as well as the techniques and strategies required in statistical modeling are described in chapters ▶Regression Methods for Epidemiological Analysis and ▶Survival Analysis of this handbook. A particular model that attempts to quantify trends in rates over time as a function of age, period of event, and year of birth is described in detail in Sect. 5.5.3.3 of this chapter.

5.4 Study Designs in Descriptive Epidemiology

5.4.1 Data on Individuals

Investigation of observed associations between exposure variables, and disease outcome in individuals, may be investigated using cohort (prospective) (cf. chapter ▶Cohort Studies of this handbook), case-control (retrospective) (cf. chapter ▶Case-Control Studies), and cross-sectional study designs.

5.4.1.1 Cohort Designs

The basic descriptive epidemiological study involves a comparison of disease rates in individuals categorized according to exposure variables that have been obtained from "routine" data sources and relate to personal (demographic) characteristics, place (of residence, birth, diagnosis), and time (of birth, of diagnosis). The focus is on the risk in one exposure group, relative to another. Relative risks may be approximated by the ratio of disease rates, with person-years at risk estimated from census data (or population registers, if available). The simplest way to adjust for the major confounders (especially age) is by standardization of the rates used for the rate ratios (Sect. 5.3.1.1). The idea of comparing summary rates for two populations seems appealing since we would hope to describe the differences between the two as a simple ratio. However, this simple description would only be appropriate if the proportional differences in age-specific rates were constant across all age groups, that is, if the assumption of proportionality holds (see Sect. 5.3.1.2). The multiplicative model presented in Sect. 5.3.1.2 is important in descriptive cancer epidemiology, since the aim is generally to estimate ρ and its statistical significance (Breslow 1984; Estève et al. 1994).

Alternatives to the rate ratios of age-standardized rates are the Mantel-Haenszel estimate (MH) introduced in Sect. 5.3.1.2, which has been shown to be particularly robust, and the internal standardization method of maximum likelihood (ML), which has optimal statistical properties (Breslow and Day 1975). Providing the assumption of proportionality of the ratio of the age-specific rates in the two groups is valid, the values obtained from the three methods should be close.

An alternative method of comparison is to calculate the standardized incidence ratio (SIR) for the populations being compared to a reference population

(Sect. 5.3.1.1 "The Indirect Method"). This indirect standardization is often preferred to direct standardization to increase statistical precision for rare diseases or small populations. As already mentioned in Sect. 5.3.2, when more than one confounding variable (age) is present, the adjustment methods discussed above are not suitable, and it is more efficient to use standard log-linear modeling methods (Kaldor et al. 1990; Kirkwood and Sterne 2003). When the population-at-risk in each cell of the cross-classification is available, it is assumed that the number of cases or deaths per cell has a Poisson distribution, with mean value proportional to the number of person-years at risk, and that the logarithm of the rate is a linear function of the classification variables. Poisson regression provides adjusted relative risk estimates for each population group, with reference to an appropriate standard. Even when age is the only confounding variable, statistical modeling has advantages over standardization and related techniques in relative risk estimates with greater numerical stability (Breslow and Day 1987).

5.4.1.2 Case-Control Comparisons

Very often, in descriptive studies, the information on the cases/deaths is more detailed, in terms of variables of interest, than that on the population at risk. For example, the case file may include information on occupation, socioeconomic status, or details about date of immigration, while the population-at-risk cannot be categorized in such detail. Analysis has to rely entirely on the numerator data, that is, on proportionate incidence or mortality data. Comparison of proportions between different case series via the proportionate mortality ratio (*PMR*) or proportionate incidence ratio (*PIR*) is generally, implicitly at least, an attempt to approximate the relative risk or ratio of rates. Confounding by age (as a result of different age structures of the case series being compared) can be removed by indirect standardization techniques, with the fraction of deaths or of cases due to specific causes in the reference series as the standard (Breslow and Day 1987). Proportionate methods are, however, relative measures, and the *PMR* for a specific disease (age standardized) is close in value to the ratio of the *SMR* for the disease, to the *SMR* for all causes (Kupper et al. 1978).

Odds ratios provide a better estimate of relative risk than the ratio of proportions, in most circumstances (Miettinen and Wang 1981). Odds ratios are estimated by case-control comparisons, comparing exposure status among the cases of the disease of interest, and cases/deaths of other (control) diseases. Unconditional logistic regression or stratified analyses are performed to obtain maximum likelihood estimates of the odds ratios (Breslow and Day 1980). The odds ratio values based on logistic regression are heavily dependent on the choice of the controls: if the risk for the subjects used as controls is unrelated to exposure, the estimates from the logistic model for the effects of exposure closely approximate those which would be obtained using Poisson regression with denominator populations.

These methods have been widely used in the study of disease risk by social status and occupation (Logan 1982) and in migrant populations (Marmot et al. 1984; Kaldor et al. 1990).

5.4.1.3 Cross-Sectional Studies

In contrast to longitudinal studies, for which observations of cause and effect represent different points in time, cross-sectional studies *simultaneously* observe exposure status and outcome status at a single point in time or over a short period in the life of members of a sample population. The analysis proceeds by determining the prevalence "rates" in exposed and non-exposed persons or, according to level of exposure, commonly using data from complete population surveys to correlate putative etiological factors with outcome.

Cross-sectional studies are relevant to public health planning, in that they may provide information on the care requirements of a population at a given point in time. In the context of investigative inquiries, they can be considered advantageous to more complex designs in measuring the association between diseases of slow inception and long duration, where the time of onset is difficult to establish, such as osteoarthritis or certain mental disorders, and exposures that endure over a prolonged time period, for example, HLA antigens or air pollution levels. They are simpler, quicker, and more economical than cohort (longitudinal studies), as no follow-up of individuals is required. In addition, the sample may be more representative of the target population as they are based on a sample of the general population.

There are however two major drawbacks to cross-sectional studies. Firstly, it is difficult to establish whether the temporal sequence is from that of exposure to outcome or vice versa. Are individuals, for instance, in lower socioeconomic groups more likely to develop mental disorders, or is it that mental illness triggers a series of events that relegates persons from a wide spectrum of socioeconomic groups to a lower status at the time of measurement? Thus, it is important to consider the possibility of reverse causation – whereby exposure status is in part a consequence of disease; the association obtained in a cross-sectional study may be wholly different from that obtained at time of the disease origin. Secondly, as it is usually not possible to determine incidence in cross-sectional studies, the use of prevalence as a proxy of frequency may distort the exposure-disease relationship as, by definition, the prevalence measure will include a larger number of cases with a long duration of disease relative to incidence. Hence, persons who die or recover quickly tend to be less likely to be included as a prevalent case than persons with long-lasting disease.

In descriptive studies, measures of exposure and outcome are taken directly from existing survey datasets. These include general purpose datasets, such as that derived from the General Lifestyle Survey in the UK (Office for National Statistics 2012), a multipurpose continuous survey which collects information on a range of topics, including health and the use of health services from people living in private households in Great Britain. Other datasets, more specific to health topics, include the National Health Interview Survey (NHIS) and National Health and Nutrition Examination Survey (NHANES) carried out by the National Center for Health Statistics (CDC 2012) in the USA.

The analysis proceeds by classifying exposure and outcome status dichotomously in a contingency table. A *prevalence rate ratio* can be calculated as the ratio of the prevalence of the outcome in those exposed to the putative risk factor compared with those not exposed or, where the level of exposure varies by intensity, test for a trend in outcome by exposure category. A case-control approach to the analysis can also be taken, whereby the *odds ratio* is calculated, although it is important to appreciate that the two ratios are not equivalent and only approximate each other when the prevalence and odds are small and the disease is rare (cross-sectional studies however require relatively common outcomes). Confounding is an important bias (cf. chapter ▸Confounding and Interaction of this handbook), and multivariate techniques such as logistic regression (estimating odds ratios) and proportional hazards models (estimating prevalence rate ratios) can be used when the outcome is, for example, the presence or absence of disease.

5.4.2 Ecological Studies

5.4.2.1 Main Characteristics

The characteristic of ecological studies is that *exposure* and *outcome* are measured on *populations/groups*, rather than on individuals. The units of observation may be populations defined by place of residence (counties, regions, districts, etc.); by personal characteristics, such as race, religion, or socioeconomic status; or by time (birth cohorts). Usually, these studies are descriptive, in that they exploit preexisting sources of information, rather than data collected to investigate a specific hypothesis. Thus, the outcome (disease) data are the likes of mortality rates, incidence rates, or prevalence data from health surveys. Exposure information may be from sources such as household/community surveys, environmental measurements, or commercial sources (data on production or sales). Exposure is expressed as an aggregate (summary) measure such as population mean, median, and proportion based on observations from individuals within the group. Exposure data may also be some environmental measurement (e.g., of air pollution, ambient temperature). The essential difference from most epidemiological study designs is that there is no information on the joint distribution of exposure and outcome in the individuals *within* the populations being studied. Table 5.3 shows two populations (A and B). In studies in which data on individuals are available, the numbers in the individual cells of each table are known, and we may calculate relative risks or odds ratios for each (and combine the results from the two strata). However, in an ecological study, only the values in the margins of each table (with n_{EA} denoting the number of exposed and $n_{\bar{E}A}$ the number of non-exposed subjects in population A and n_{DA} denoting the number of diseased and $n_{\bar{D}A}$ the number of non-diseased subjects in population A, analogously for population B) are known, so that we have simply prevalence of exposure, and disease incidence or prevalence, for populations A and B.

Ecological studies may be used to generate (or test) etiological hypotheses and to evaluate interventions at the population level.

The main problem, as described below, is that ecological designs are usually being employed to make such inferences (concerning cause/prevention) about individuals, based upon observations using groups. Such interpretations are prey to

Table 5.3 Ecological study: comparison of prevalence of exposure and outcome in two populations A and B

	A				B		
	Diseased	Non-diseased	Total		Diseased	Non-diseased	Total
Exposed	?	?	n_{EA}	Exposed	?	?	n_{EB}
Non-exposed	?	?	$n_{\bar{E}A}$	Non-exposed	?	?	$n_{\bar{E}B}$
Total	n_{DA}	$n_{\bar{D}A}$		Total	n_{DB}	$n_{\bar{D}B}$	

Prevalence of exposure in A = $\frac{n_{EA}}{n_{EA}+n_{\bar{E}A}}$

Prevalence of exposure in B = $\frac{n_{EB}}{n_{EB}+n_{\bar{E}B}}$

Rate of disease in A = $\frac{n_{DA}}{n_{DA}+n_{\bar{D}A}}$

Rate of disease in B = $\frac{n_{DB}}{n_{DB}+n_{\bar{D}B}}$

a variety of artifacts, referred to collectively as “the ecological fallacy” (Piantadosi et al. 1988). However, ecological studies may have particular value when some characteristic of the group (rather than the sum of individuals within it) is important in determining outcome, so that ecological designs are more appropriate than studies of individuals within the groups (see Sect. 5.4.2.4). Similarly, ecological studies may be used to evaluate the effect of population-level interventions, especially if the interest is in the effect at group rather than individual level. Thus, the relationship between exercise and mortality from cardiovascular disease may be known from individual-based studies, but the effectiveness of an educational program on the topic in influencing disease rates might choose an ecological design to capture the combined effect at group and individual level.

5.4.2.2 Types of Design

Exploratory Ecological Studies The term “exploratory ecological study” has sometimes been used to describe the comparison of disease rates between populations defined, for example, by place of residence, ethnic group, birthplace, or birth cohort (Estève et al. 1994; Morgenstern 1982). In fact, it is hard to understand the justification for the use of the term “ecological” for such studies. They compare disease risk among individuals characterized by various exposure variables (such as place of residence, or birthplace, or period of birth) and as such, differ only in the source of information, from cohort studies using questionnaires or biological measurements, comparing disease rates according to exposure type or level (occupational groups, smoking status, etc.). Probably, the term should be reserved for comparisons between groups in which “exposure” has not been measured but is simply assumed from some sort of a priori knowledge or guesswork. This is the basis of many studies carried out with quite sophisticated laboratory methods, where the subjects comprise some sort of sample (usually by no means random) from populations believed to be at high/medium/low exposure of something. Another version would be studies in which exposure is not measured, and may not even be defined, but is assumed to have some underlying spatial or temporal distribution; the purpose of the analysis is to see if disease risk in the population groups studied has,

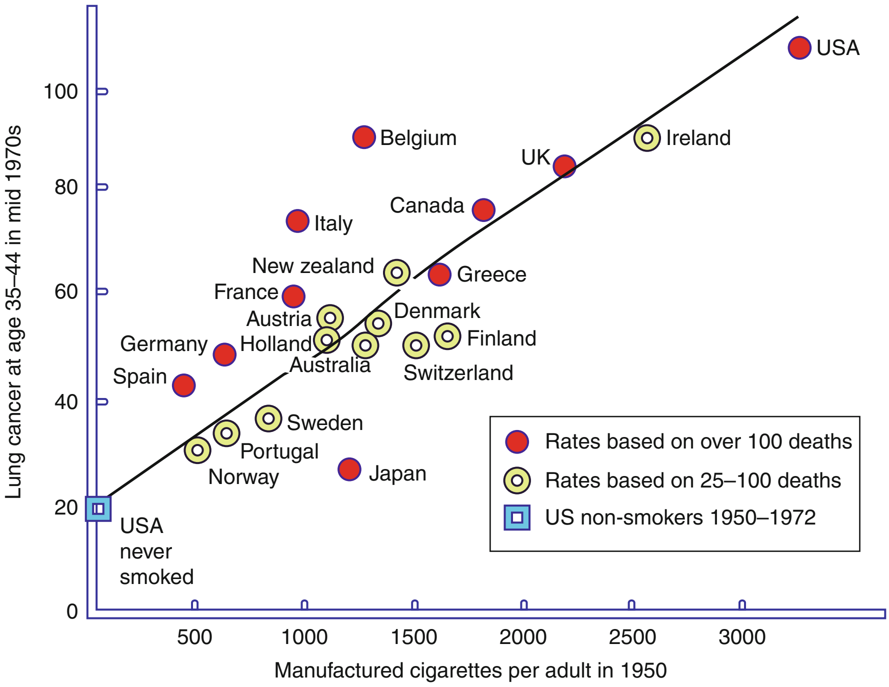


Fig. 5.3 International correlation between manufactured cigarette consumption per adult in 1950 while one particular generation was entering adult life (in 1950), and lung cancer rates in that generation enters middle age (in the mid-1970s)

too. This includes studies of geographical clustering and of spatial autocorrelation (cf. chapter ►Geographical Epidemiology of this handbook).

Multigroup Comparison Ecological Study This is the most commonly used study design. For several populations (usually geographical regions), outcome (disease) levels (prevalence, rates) are compared with exposure (means, proportions) variables of interest. An example is given by the early studies suggesting the importance of blood lipids in the etiology of ischemic heart disease. Coronary heart disease rates were compared with plasma cholesterol and dietary fat intake in different populations (McGill 1968). Figure 5.3 shows a well-known example from Doll and Peto (1981), relating lung cancer mortality to consumption of manufactured cigarettes.

Time Trend Ecological Studies A single population is studied, but is cut up into *groups* corresponding to different time periods. The objective of the study is to determine whether the time trend in outcome (disease rates) corresponds to time trend in exposure.

Multiple Group Time-Trend Ecological Studies These are a mixture of multi-group and time trend designs. Change in exposure and outcome over time is compared for several populations. Dwyer and Hetzel (1980) compared time trends in coronary heart disease mortality in three countries in relation to changes in major risk factors. The advantage of this design is that it is less subject to confounding than with a single population, that is, the unmeasured factor, related to both exposure and outcome (change in disease), is unlikely in several different populations but quite possible in one.

5.4.2.3 Analytical Methods

The simplest level of analysis is to plot the disease rate and indicator of exposure for each population on a scattergram and to calculate a correlation coefficient. This merely indicates the level (strength and direction) of association between the parameters; it does not necessarily imply that the exposure variables predict outcome, rather that other influencing factors (confounders) are likely to have been well controlled in the ecological grouping. Moreover, the correlation coefficients may be quite biased, especially if the groups for study have been chosen on the basis of their level of exposure.

More usually, interest lies in quantifying the magnitude of the effect to be expected from different levels of exposure, and regression of group-specific disease rates (Y) on group-specific exposure prevalence (x) is the method employed. The simple linear model ($Y = \alpha + \beta x + \varepsilon$ with ε denoting an error term) is typically used. An estimate of the effect of exposure (at the individual level) can be derived from the regression results (Beral et al. 1979). The relative risk is the ratio of the disease rate (Y) in an exposed population ($x = 1$) divided by the rate in an non-exposed population ($x = 0$). Assuming the above linear model for Y, this results in

$$RR = \frac{\alpha + \beta * 1}{\alpha + \beta * 0} = \frac{\alpha + \beta}{\alpha} = 1 + \frac{\beta}{\alpha}.$$

If a log-linear model is fitted, such that $\ln(Y) = \alpha + \beta x + \varepsilon$, then the estimate of relative risk can be derived as $\exp(\hat{\beta})$. For more details on regression models, we refer to chapter ▶Regression Methods for Epidemiological Analysis of this handbook.

These equations assume that the groups studied are perfectly homogenous for exposure and that the relationship modeled (prediction of disease rate) is valid at both extremes of exposure (nil or total), a situation that is rarely observed in practice. Homogeneity of exposure is unlikely, of course, and the summary statistics (means, medians) have large and unknown error terms. Trying to mitigate the problem by studying small population gives rise to different technical problems (measurement error, migration) and a larger variance of estimated disease rates.

In the situation where the rates in the different populations being compared have different precision (due to varying size), weighted regression is frequently used, to give more emphasis to the larger units. The usual weighting applied is the inverse of the variance, although maximum likelihood methods (taking into account variation in rates that would be expected by chance) may be more appropriate (Pocock et al. 1981).

Time Lagging It is reasonable to take exposure date from an earlier period than that for the outcome (disease). Varying the interval to obtain the best fit (e.g., correlation) has been used to provide information on the possible induction period between exposure and disease. Rose (1982) found that the correlation between serum cholesterol and coronary heart disease mortality in men aged 40–59 in seven countries was maximal when the interval between the two measures was 15 years.

5.4.2.4 Advantages of Ecological Studies

There are several advantages to ecological study designs:

- They are very economical, since they use existing data on exposure and outcome, with no costs involved in collection.
- They are very rapid, even compared with case-control studies, where time is needed for recruitment, for example, for investigating suspect clusters.
- Very large numbers can be studied, so that *small* increases in risk can be investigated. Small risks affecting large numbers of people are important from a public health point of view.
- They may – and ideally do – include populations with a very wide range of exposure level (more than can be found in a single population used for conventional cohort or case-control studies). For example, the range of variation in dietary fat intake in a single population may be too small to demonstrate the differences in risk at different levels (Prentice and Sheppard 1990).
- They may be the only practical analytical approach to investigating the effects of an exposure that is *relatively constant* in a population but differs *between* populations, for example, exposure to external environment (air, radiation, water).
- In many circumstances, individual measures of exposure are difficult or impossible to obtain. This is often the case in studies of diet and disease, since collection of individual food records is difficult, may not reflect habitual intake, and does not allow for individual variability in metabolic response to a given diet. 24-hour dietary recalls, although of little value in the study of dietary exposures in individuals, may, when averaged for a population, provide a useful indicator of exposure for ecological analysis. The same principles apply to average exposures to air pollution, trace elements in soil/water, and so on.
- There may be interest in "contextual effects," that is, group or community level effects, rather than inferences about individual exposure-outcome. This is particularly important in communicable disease epidemiology, where models of transmission have a group component, for example, transmission of vector-borne diseases will be dependent upon the prevalence of carriers in the population (Koopman and Longini 1994). Ecological studies are relevant in examining the effects of policy, laws, and social processes where contextual as well as individual effects are relevant to the outcome.

5.4.2.5 Disadvantages of Ecological Studies

The above advantages have to be contrasted with several disadvantages, not only from a technical perspective.

Technical Disadvantages

- *Data Problems*

  The data on exposure are obtained from existing sources, usually not compiled for the purpose for which they are being used, and may in consequence lead to a somewhat inaccurate estimator of the relevant exposure. Thus, data based on production, sales, or food disappearance give only an approximate guide to actual exposure, even at the group level. If the data are from a sample of the population, this may be unrepresentative of the group. The outcome (disease) variables are subject to similar concerns. Data quality may differ between the populations, due, for example, to varying completeness of death registration. Lack of comparability may also result from changes in disease classification or coding over time. A further concern is the accuracy of person-years at risk for group data. This is an issue for all types of descriptive study, where person-years at risk are estimated from cross-sectional counts, rather than longitudinal observation of individuals. Moreover, migration of populations will make comparisons over time difficult since cases of disease may not be exactly from the exposed population. This problem is compounded if migration is related to the presence of the disease studied. A solution commonly adopted is simply to assert that the populations studied are "stable," without having any objective evidence to that effect.

  Misclassification of exposure within the groups being studied may have surprising consequences. Even when non-differential (unrelated to outcome), the bias in estimated risk of exposure may be away from the null value – the opposite to the familiar situation in individual-based studies (Brenner et al. 1992).
- *Availability of Data*

  The number of variables available for the populations studied is quite limited. There is little scope for any adjustment for possible confounding (limited though this is in ecological studies).
- *Problem of Induction Period*

  It is reasonable to assume that there is some delay between exposure and outcome and that both should not be measured at the same time. It is not difficult to find published studies in which the exposure measures postdate those of outcome. Some time lag should be included (e.g., Fig. 5.3), but this involves assumptions, possibly arbitrary, as to the appropriate mean interval to use. Furthermore, exposure information may not be available for the relatively distant past, and, when it is, the population in the units of analysis will comprise different individuals, if exposure and outcome measures are far apart in time.
- *Ecological Fallacy*

  This is the main problem when inferences about individual exposure-outcome associations are being inferred from observations at the group level (Piantadosi et al. 1988). In this instance, the assumption being made is that the single measure

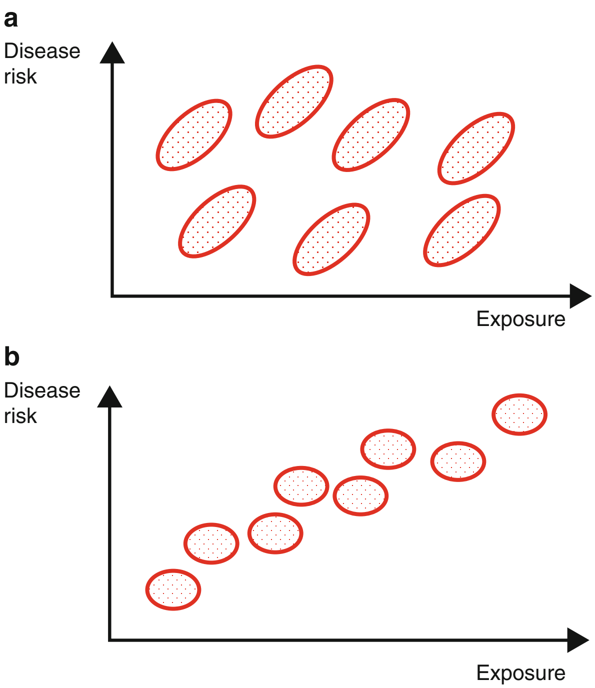


Fig. 5.4 Two situations illustrating the difference between ecological and individual associations between exposure and disease (From Walter 1991)

of exposure applies to all members of the group. This is rarely so. Often, it is obvious (e.g., exposure variable is a group mean, with its own variance). Else, it is intuitive (e.g., exposure variable is an environmental measure – e.g., solar UV, nitrate in drinking water), but individual behaviors result in varying exposure to it (wearing hats, long holidays or work periods elsewhere, using bottled water, etc.). As a result, there may be a difference between the relationship at the individual level (within groups) and group level (between groups). Ecological associations may be weaker, or stronger, than relevant associations at individual region. Figure 5.4 (from Walter 1991) illustrates two extreme scenarios. In A, there is a strong covariance between exposure and outcome within groups, but a very weak one between groups. An ecological analysis, based on a single aggregate measure of exposure and outcome for each group would show only a weak association. In B, the association within groups is weak but appears strong when examined between groups.

- *Ecological Bias*
  Due to the failure of the expected ecological effect estimates to reflect the biological effect at the individual level, two forms of bias are said to exist (Morgenstern 1982). Aggregation bias: data are aggregated, ignoring the information from the subgroups from which the individual observations came. Specification bias: a problem of using groups that in some way are related to the disease (irrespective of the exposure under study). It may result from extraneous risk factors being differentially distributed by group or from property of the

Table 5.4 An example of ecological (cross-level) bias

	Prevalence	Exposed			Non-exposed			Total		
Population	of exposure	n_E	N_E	R_E	$n_{\bar{E}}$	$N_{\bar{E}}$	$R_{\bar{E}}$	n_T	N_T	R_T
A	0.25	1,200	1×10^6	1.2	1,800	3×10^6	0.6	3,000	4×10^6	0.75
B	0.50	3,330	2×10^6	1.67	1,670	2×10^6	0.84	5,000	4×10^6	1.25
C	0.75	6,000	3×10^6	2.0	1,000	1×10^6	1.0	7,000	4×10^6	1.75
All	(0.50)	10,530	6×10^6	1.76	4,470	6×10^6	0.75	15,000	12×10^6	1.25

group itself (contextual effects). The sum of these two components provides "the ecological bias" (cross-level bias) which is present. Table 5.4 gives an example.

The relative risk of exposure in each of the three populations, A, B, and C, is 2.0 ($R_E/R_{\bar{E}}$ with R_E and $R_{\bar{E}}$ denoting the disease rates in the exposed and non-exposed subjects, respectively), that is, there is no difference in the effect of exposure within the different groups. Although the overall (crude) relative risk, summing the cases and populations at risk for the three groups, is 2.35 = 1.76/0.75, the relative risk, standardized for group, can be estimated by

$$\widehat{RR} = \frac{\sum n_E}{\sum (n_{\bar{E}} N_E)/N_{\bar{E}}},$$

where the summation is across groups, with n_E denoting the number of diseased subjects among the exposed and N_E denoting the total number of exposed subjects in the population (analogously for the non-exposed and the total population). In the example shown, therefore

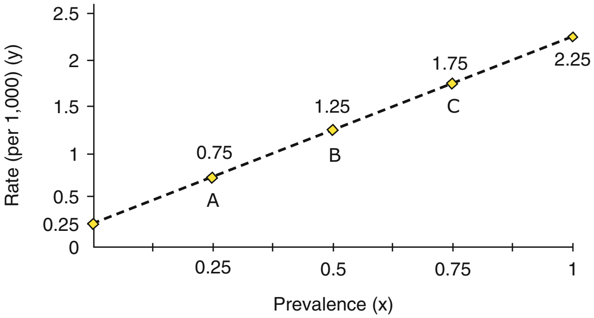
$$\widehat{RR} = \frac{1{,}200 + 3{,}330 + 6{,}000}{(1{,}800 * 1)/3 + (1{,}670 * 2)/2 + (1{,}000 * 3)/1} = 2.0.$$

The difference between the unadjusted and adjusted estimates (2.35 and 2.0) shows that there is confounding by other risk factors, which are different between the groups (as shown by different risks in the non-exposed) indicating a specification bias. Figure 5.5 shows the linear regression of the rate of disease (R_T) on the prevalence of exposure and the ecological estimate of relative risk (Sect. 5.4.2.3). The large difference between the estimates based on ecological and crude individual data (9.0 and 2.35) is the result of aggregation bias since the extraneous factor that increases the risk among the non-exposed is most prevalent in the most exposed population (C).

Confounding and Effect Modification in Ecological Studies

Confounding in epidemiological studies arises when two exposure variables are statistically associated (correlated), and at least one of them is also an independent risk factor for the disease under study, so that both will appear to be so if examined separately. In individual-based studies, with many subjects, it is feasible to separate their effects (by stratification or multivariate methods), because perfect correlation

Fig. 5.5 Ecological analysis (linear regression: $y = \alpha + \beta x$) of the hypothetical data summarized in Table 5.4: $\alpha = 0.25$, $\beta = y - \alpha / x = 2.0$, $RR = 1 + \beta / \alpha = 9.0$



between variables is very unlikely. When groups are studied, however, there may be perfect correlation between variables, particularly if the populations studied are few in number (e.g., the hi-lo two group studies beloved of laboratory workers) and large in size.

Furthermore, risk factors that are independent of exposure at the individual level may be correlated with it, and thus be confounders, when aggregated at the population level. Conversely, a confounding variable at the individual level may not be so at the ecological level; for example, although the risk of most cancers is quite different in males and females, sex, as an ecological variable, will not be associated with disease rates in geographical areas because the ratio of males to females is broadly similar in all.

Effect modification (or interaction) refers to variation in the magnitude of the effect of an exposure across the levels of a third (covariate). Effect modification can be present in an ecological association, even when not evident at the individual level. Greenland and Morgenstern (1989) give an example of a cofactor (e.g., nutritional deficiency) with different prevalence in the populations (regions) studied, which is not a risk factor in the absence of the study factor (smoking). Thus, the non-smoker rates would be the same in all regions (and region would not, therefore, be a confounder in an individual-based study of smoking and disease), but the effect of smoking would differ by region.

5.5 Descriptive Studies

5.5.1 Personal Characteristics

Routine sources of information on morbidity and mortality include information on the so-called "demographic" variables (age, sex, marital status, religion, race, education, occupation, etc.) of the cases, and the corresponding data on population-at-risk may likewise be available. This allows investigations of how characteristics of the individual relate to the risk of disease. Since they may often have striking effects on disease intensity, a number of these variables (particularly age) can be considered as among the foremost risk factors for many diseases. Exploration of the relationship between personal characteristics and disease has generated and confirmed many hypotheses and, importantly, elucidated particular mechanisms concerning other putative factors, by taking account of the strong confounding effects of routine variables that may otherwise have distorted the relationship between outcome and the exposure of interest.

5.5.1.1 Age

The increases with age in morbidity of, and mortality from, disease are more apparent than for any other variable. Excluding accidental and violent deaths, there is a 500-fold variation in the death rate from all causes between the ages of 20 and 80 (Peto and Doll 1997). For epithelial cancers, as well as for cardiovascular disease and chronic respiratory disease, there is more than a 1,000-fold difference. The age-specific patterns also differ between and within diseases; Fig. 5.6 compares the age-specific incidence of several types of cancer. The effects of age are most commonly ascribed to an individual's cumulative exposure to environmental insults (e.g., sociocultural or behavioral factors) over a life span or, in the case of, for example, breast cancer, to the effects of hormonal changes. While the process of aging is commonly put forward as a possible mechanism in its own right, for example, through declining immunological defenses, or an increasing number of mutations in certain somatic cells (Lilienfeld and Lilienfeld 1980), others suggest that while aging is clearly related to disease, there is no evidence that aging itself is a biological process that causes disease (Peto and Doll 1997).

The fundamental importance of age as a major confounder in almost all epidemiological studies is exemplified by age standardization or stratification to control for its effects (see Sect. 5.3.1.1).

Interesting forms of the distribution of disease risk by age have motivated a series of hypotheses as to the biological mechanisms underlying particular diseases. The bimodality of Hodgkin lymphoma (Fig. 5.6) suggested that it comprised at least two distinct forms of the cancer and the likelihood of differing etiologies. Early investigations seeking biological explanations for particular age-disease patterns led to hypotheses concerning the importance of early development to disease later in life. Recently, frailty models have been used to examine heterogeneity in incidence and obtain biologically relevant estimates of the proportion of susceptible individuals and the number of events required for cancers to develop (e.g., Moger et al. 2009). More recently, a life course approach to chronic and infectious disease has been conceptualized, which considers the long-term effects of factors during gestation, childhood, and adolescence on subsequent adult morbidity or mortality (Kuh and Ben-Shlomo 1997; cf. chapter ▸Life Course Epidemiology of this handbook).

The importance of period of birth (birth cohort) is clear when investigating changes in disease risk over time (see Sect. 5.5.3.3); it was through the study of age curves that the influence of generation effects on disease was first realized, however. In examining the age-specific mortality rates from tuberculosis in different calendar periods of time, Frost (1939) after Andvord (1930) showed that the peak in more recent cross-sectional age-mortality curves (in 1930) at later ages (50–60) compared to peaks in young children (0–4) previously (in 1880) and at the ages 20–40 (in 1910) was an illusion – an examination of the same age curves by cohort indicated subjects comprising the 1930 age curve passed through greater risks in previous

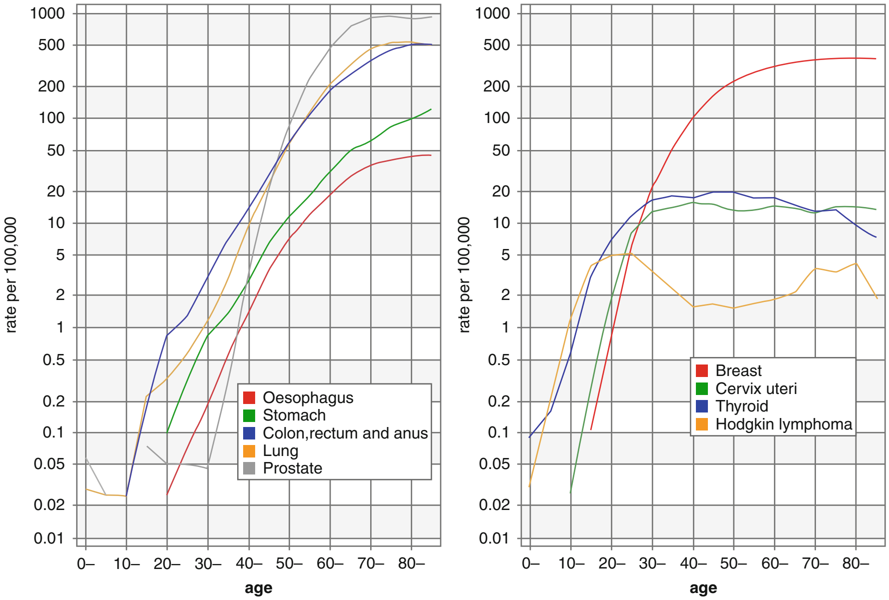


Fig. 5.6 Age-specific incidence rates. Canada, selected cancer sites, 1998–2002

decades – the class of individuals who were children in 1880 and who were aged 50–60 (if still alive) by 1930 (Fig. 5.7). In concluding, he noted that contemporary peaks of mortality in later life did not signify a postponement of maximum risk but rather were the residuals of higher rates in early life.

Korteweg (1951) similarly demonstrated that a cross-sectional view of age curves of lung cancer mortality led to an erroneous interpretation as the age curves were artificially pushed down by the increase in lung cancer in younger age groups – the consistent pattern of declining rates at relatively early ages (65 and over) for five consecutive periods between 1911 and 1945 was therefore not an observation that required a biological explanation. The mechanisms that promoted lung cancer, he observed, acted particularly (but not exclusively) in younger people.

Several well-known biases may distort the underlying age-disease relationship. The quality of mortality statistics in the very elderly is particularly affected by the precision and coding of the death certificate, as well as the decision as to the underlying cause of death. For incidence data, case ascertainment is less effective in the very old, in part due to inaccuracy in the abstraction and coding of diagnostic information and in part due to competing causes of death.

5.5.1.2 Sex

There are, for certain diseases, substantial differentials in the rates in men compared to women (sex ratio) that may represent fundamental differences in exposure to environmental risk factors and/or response to them. Mortality rates from several common causes of death, such as ischemic heart disease, malignant neoplasms,

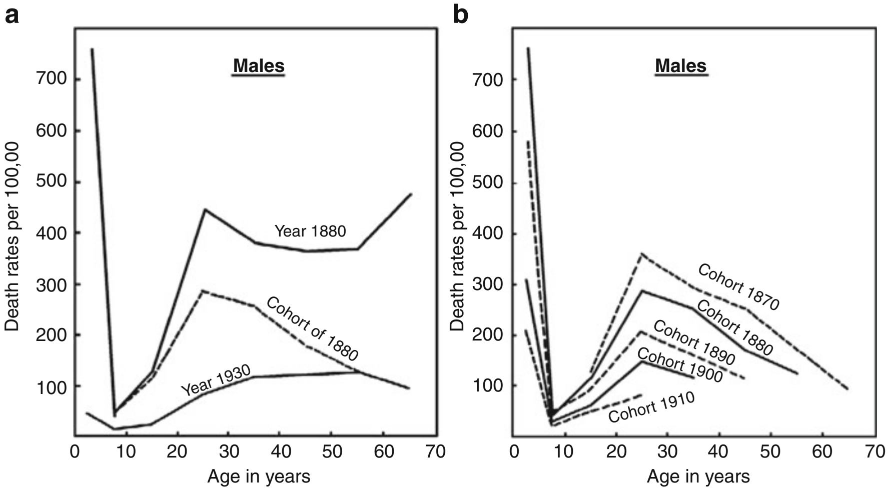


Fig. 5.7 Age-specific mortality rates from tuberculosis in Massachusetts, by period (**a**) and by birth cohort (**b**) (Frost 1939)

and HIV-AIDS (in Western countries), have sex ratios substantially greater than one. The disparity is perhaps not surprising given the contrasting social, cultural, and behavioral practices of men and women and the strong lifestyle component of such diseases. Other than environmental exposures, endogenous factors, such as the sex hormones, may contribute to differences in risk between the sexes, acting as promoters of disease pathogenesis or as protective factors, while our understanding of the putative impact of sex-specific genetic predisposing factors is in its infancy.

The marked differences between the sexes in the incidence of some common cancers are shown in Fig. 5.8; for many of these neoplasms, much of the variation can be explained by contrasting levels of exposure to well-established carcinogens in men relative to women. Hence, the high male:female (M:F ratio) for mesothelioma is largely a consequence of historical exposure to asbestos in men through certain occupations, while lung cancer largely replicates the past history of tobacco smoking in men relative to women. Based on site-specific M:F ratios of age-standardized rates in developed countries, the majority of the cancers of the head and neck, as well as of the bladder and esophagus, are also much more common in men reflecting the heavier alcohol consumption acting independently and multiplicatively with tobacco smoking.

For M:F ratios lower than one, the most outstanding example is for breast cancer for which there is a 500-fold difference in risk in women relative men, which might be attributable to the mammary gland mass, as a correlate of the number of cells susceptible to transformation, as well as hormonal milieu. Differences in gallbladder cancer are probably attributable to a higher prevalence of gallstones in women relative to men. A corresponding distribution of sex ratios is observed for cancer mortality, but additionally, differences in survival (through gender rather than biological differences, e.g., stage of presentation) modify the differentials.

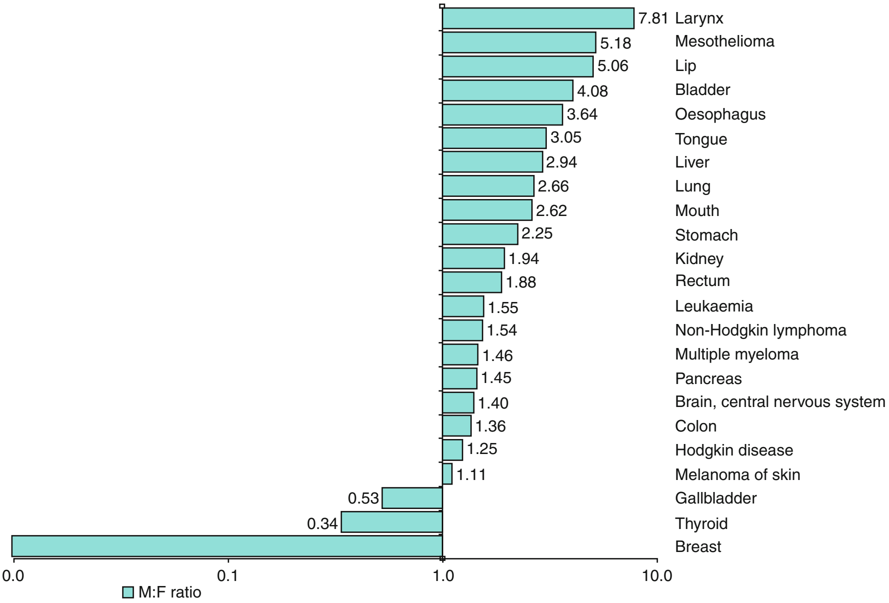


Fig. 5.8 M:F ratios of age-standardized rates for the common cancers in developed areas worldwide (Parkin et al. 2002)

5.5.1.3 Ethnic Group

Variations in the risk of disease and differences in the health experience of individuals from different ethnic groups have been the subject of many studies (Macbeth and Shetty 2001). Studies within multiethnic societies are more valuable than international comparisons, if the primary variable of interest is ethnicity or racial group, since at least some of the environmental differences present in international comparisons are reduced or eliminated. There are plenty of examples of such studies from multiethnic populations in all parts of the world.

Interpretation of ethnic differences in risk should first consider the possibility of data artifact. Even within a single country, it is possible that differential access to health-care and diagnostic services by ethnic group may influence reporting rates of disease, for example, access to and acceptance of screening programs has been shown to differ by ethnic group in several countries (Parker et al. 1998; Seow et al. 1997). Differences in access to treatment certainly can affect outcome, so that survival rates from cancer are well known to vary by race/ethnicity (Baquet and Ringen 1986); since mortality rates are determined by both incidence of disease and survival, this is a major consideration if mortality is being used, as it often is, to provide information on cancer risk.

Artifact aside, the principal question posed by observed interethnic differences in risk is how much is due to variation in exposure (to “carcinogens” or “risk factors”),

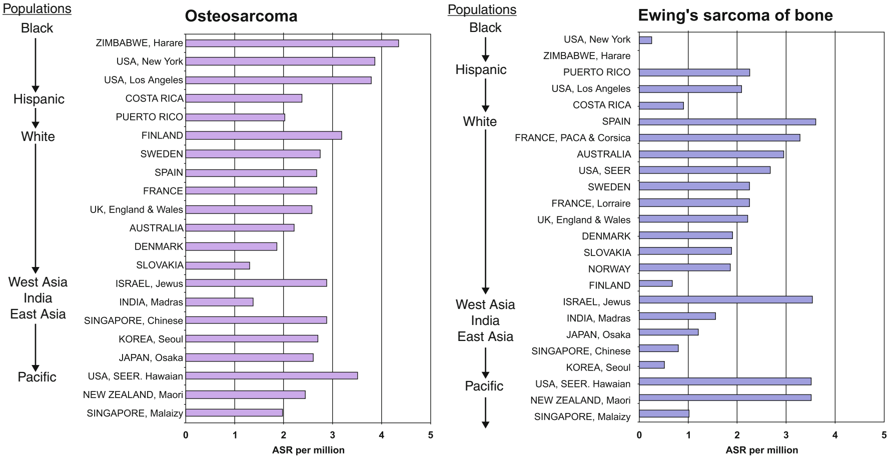


Fig. 5.9 Incidence of osteosarcoma and Ewing's sarcoma in childhood (Parkin et al. 1998)

and how much is the result of inherent differences in susceptibility to such exposures (and hence genetically determined).

From an epidemiological point of view, the variable "ethnicity" or "race" defines a constellation of genetic factors, which relate to susceptibility to a given disease. Of course, there is considerable variation *within* a given ethnic or racial group (however this is defined), but there are often sufficiently large differences between them to yield distinctive patterns of risk. If "ethnicity" is the variable of interest, the first consideration is to eliminate the effect of confounding variables, associated with the risk of disease and differentially distributed by ethnic group. The relevant exposure variable is quite likely to be exposures such as tobacco, alcohol, diet, and infection, but in descriptive studies, there will generally only be information on so-called demographic variables such as social class (or occupation, educational level), place of residence, and marital status that are proxies for these.

A striking illustration of the likely influence of genetic factors on risk of disease is provided by certain cancers of childhood. For bone tumors (Fig. 5.9), there are very marked differences in incidence between ethnic groups for which no plausible environmental "exposure" can be imagined. Any such exposure would have to be very carcinogenic (to act so early in life), very tissue specific, and be very unevenly distributed by ethnic group.

The most fruitful approach using routine data sources is through the study of migrants (see Sect. 5.5.4.2) that attempts to separate the "genetic" and "environmental" components of differences by studying disease risk in a given migrant population in comparison with that in the host population (similar environment, different genetics) and in the population living in place of origin (similar genetics, different environment).

5.5.1.4 Socioeconomic Status

Socioeconomic status is an extremely important but rather vague term for a whole host of factors that require individual consideration and action, such as income, occupation, living conditions, education, and access to services. For many diseases, such as cardiovascular disease (Rose and Marmot 1981) and cancer (Smith et al. 1991; Kogevinas et al. 1997), a clear gradient by social class is observed, with the highest disease rates or the poorest outcome often observed within the lowest socioeconomic grouping. The influence of social status has a marked effect on disease outcomes in adults and in both prenatal development and infant mortality.

While the magnitude of gradients of many disease outcomes tend, to vary additionally with time reflecting in part changing social and economic circumstances (Marmot 1999), the impact of general improvements in health often fail, to reach the most disadvantaged: Since the 1920s, for instance, improving infant survival rates over half a century in the UK was not observed among those considered least advantaged (with the consistently highest infant mortality rates) (Rosen 1993).

A recent advance has been the linking of "deprivation" scores, an index that combines a number of social variables from censuses according to area of residence, to data records from routine data sources (e.g., cancer registries) at the small area level (Carstairs 1995). In measuring socioeconomic status, a number of potential surrogates may provide a reproducible definition, such as affluence (income), living conditions, or occupation. From a perspective of routine systems based on populations, some of the most illustrative stem from the pioneering series of reports published by the Office for National Statistics, formerly the Registrar General for England and Wales. The Registrar General's social classes were derived from a classification of occupations according to status and level of responsibility (and for married women, on the basis of their husband's employment). Figure 5.10 demonstrates the uniform socioeconomic gradients of infant mortality by cause, with the most abrupt increases observed for accidents and respiratory disease.

There are interpretational difficulties with socioeconomic data from routine sources. Although selection bias should be minimal if the comparison of health events with survey data is made from the same population at the same time, measurement bias is of particular concern. In occupational studies, the respondent is asked for only one occupation, even though they may have had a history of different professions (see Sect. 5.5.1.6).

The consistent findings of poorer disease outcomes among the more disadvantaged social groups have political as well as public health implications. In addition to pinpointing the need for health promotion and disease prevention strategies targeted at low socioeconomic groups, health inequality also demonstrates a need for change at the societal level: improving health by improving incomes, basic housing, and working conditions. However, the practical importance of the variable as an independent risk factor in epidemiological studies is perhaps overstated. Health-deprivation gradients are now well established in most disease domains, and specific public health actions are difficult to formulate and implement given social

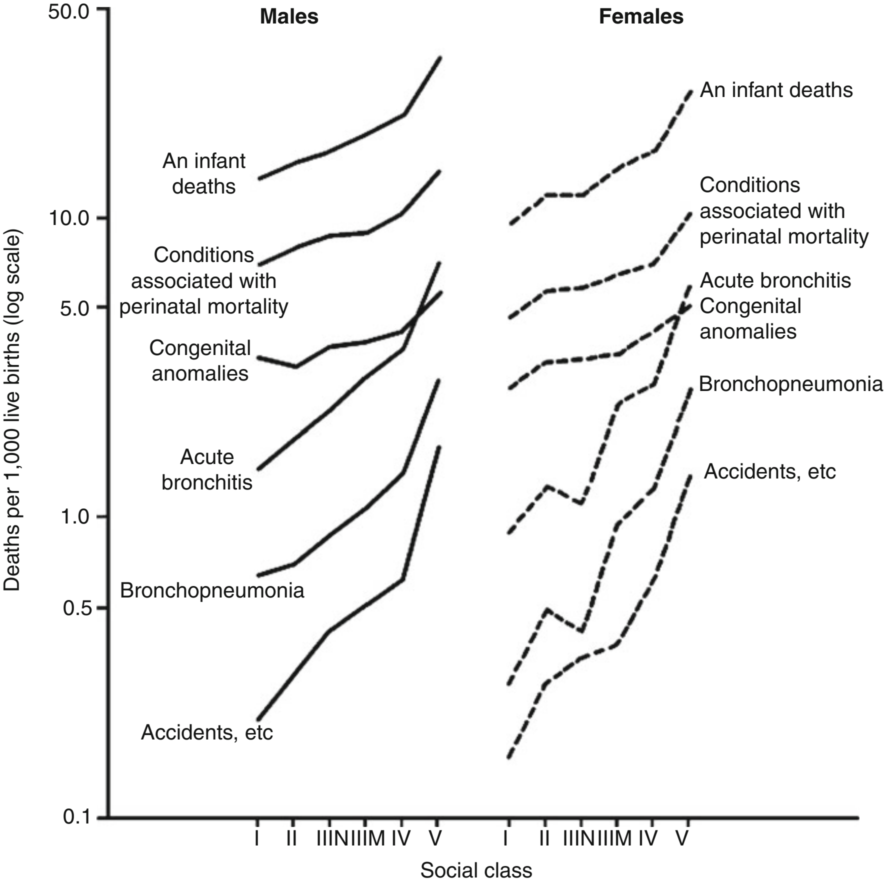


Fig. 5.10 Infant mortality by sex, occupational class, and cause of death (Occupational Mortality 1970–1972; Registrar General's Office for England and Wales 1978, p. 158)

class acts as a surrogate for a vast and complex array of social and environmental processes, congenital characteristics, and early life experiences.

5.5.1.5 Marital Status

A number of epidemiological studies have examined marital status in relation to disease risk. While it is apparent that there are major differences in disease rates according to this variable – married persons often have lower death rates than single persons – there are commonly difficulties inferring the true nature of the association. The main difficultly arises in determining whether being married per se offers health advantages or if there are certain characteristics of good health or long life that favor an individual's predisposition to marriage. The variable has, however, proved useful in determining certain surrogate populations – such as never-married males as proxies for homosexual men (Biggar et al. 1987). In this context, the variable was used to establish increases from the mid-1970s in AIDS-related cancers such as Kaposi's sarcoma in single men aged 20–49 years old (Biggar et al. 1987).

5.5.1.6 Occupation

Descriptive studies have made an important contribution to occupational epidemiology, through the analysis of routine record sources. In some countries, occupations are recorded on death certificates or in disease registers. Such routine record datasets can be used to calculate cancer risks in different occupations. If the comparisons are to involve rates (of mortality, or incidence), then suitable population-at-risk data must be available from the census or population register, with occupations classified in the same way; failing this, proportionate methods may be used (Sect. 5.4.1.2). Routine record studies are relatively inexpensive and often entail very large numbers of subjects. Occupation is specified in terms of job titles. A limiting factor of such analyses is the validity of the job title information collected. With an interviewer-administered questionnaire, quite valid job histories can be obtained, but the validity of occupations recorded on routine records such as death certificates or tumor registers is typically mediocre (Wigle et al. 1982; Steenland and Beaumont 1984; Armstrong et al. 1994). In addition, routine records typically contain only one of the subject's jobs, usually the most recent which may include "retired." In a large sample of Montreal workers, it was estimated that, on average, about 62% of working years were spent in the job of longest duration and about 50% in the last job (Siemiatycki 1996). Some of the defects of the limited information on occupation available in routine records can be reduced by linkage with more valid sources of occupational data. Examples are in Canada (Howe and Lindsay 1983) where a government-run labor force survey was linked to mortality records and especially in the Nordic countries where census data have been linked to the cancer registry (Lynge and Thygesen 1990; Andersen et al. 1999).

Despite these limitations, analyses using job titles are useful. Several associations with cancer have been discovered by means of analyses based on job titles. Table 5.5 shows results from the decennial occupational mortality analysis (Registrar General 1978) based on deaths occurring in England and Wales in 1970–1972, and cancer registrations in 1968–1969. Analyses such as these are most valid and valuable when the workers have a relatively homogeneous exposure profile – for example, miners, motor vehicle drivers, butchers, and cabinetmakers. However, job titles are limited as descriptors of occupational exposures (Siemiatycki et al. 1981). On the one hand, many job titles cover workers with very diverse exposure profiles, while on the other, multiple exposures are found to occur in many occupation categories. Several approaches have been used to better define actual exposures. One such is the job-exposure matrix (JEM). A JEM is simply an automatic set of indicators showing which exposures may occur in which occupations (Hoar et al. 1980; Siemiatycki 1996; chapter ▶Exposure Assessment of this handbook).

Table 5.5 Stomach cancer deaths and registrations by occupation unit and cancer: men aged 15–74 giving units with significantly raised proportionate mortality ratios *PMR*s and proportionate incidence ratios *PIR*s in 1968–1969 ($p < 0.01$) (Registrar General's Office for England and Wales 1978)

Cancer (ICD number)			Deaths 1970–1972			Registrations 1968–1969	
Occupation			All cancer *PMR*			All cancer *PIR*	
Order	Unit	Title	Observed			Observed	
			15–74	15–64	65–74	15–74	15–74
Stomach (151)							
II	007	Coal mine – face workers	615	142**	127**	44	182**
II		Coal miner – other underground				45	214**
II	008	Coal mine-workers above ground	120	159**	139*	34	207**
II	007	Coal miners (so described)	615	142**	127**	270	146**
IV	015	Fumacemen, kilnmen, glass, and ceramics	16	122	72	9	173
VII	054	Other metal making, working; jewelry and electrical production process workers	262	121*	133**	96	145**
X	064	Fiber preparers	35	191 *	110	13	194*
XIV	089	Workers in rubber	44	123	122	27	175**
XV	098	Construction workers nec	256	102	106	127	132**
XVI	102	Boiler firemen	121	107	120	69	143**
XVIII	106	Railway lengthmen	65	103	113	33	156*
XVIII	113	Laborers and unskilled Workers nec, building and contracting	149	109	78	56	143*
XVIII	114	Laborers and unskilled workers nec, other	1,120	114**	113**	781	122**
XIX	123	Inspectors, supervisors, transport	84	94	109	50	149**
XX	136	Warehousemen, storekeepers and assistants	664	103	111	314	129**
XXII	144	Shop salesmen and assistants	109	89	100	11	111

Those results significant at the 1% level are denoted ** and those at the 5% level*
nec not elsewhere classified

5.5.2 Place of Residence

Place of residence is an important variable in descriptive epidemiology. It is almost always available in routine sources of events of disease (registers, surveys, death certificates), and population-at-risk is very often available for small geographical units too. “Geographical pathology” – comparisons of disease rates or risk of individuals living in different areas – has been one of the longest established and productive types of descriptive study for more than a century (Hirsch 1883). National populations have often been the unit of study. The reason is that this dimension is the one for which statistics – especially mortality – are collected and published. Differences in disease between countries may indeed be striking. Sometimes, the reasons are obvious and correspond to the known distribution of causal agents – as for some infectious diseases. But for some diseases, the clear international variations in incidence or mortality have prompted research to better understand the reasons behind them.

Valid comparisons of data deriving from routine data sources in international studies require that there is fair comparability in diagnostic criteria and in recording and coding the events concerned. Variation in quality and completeness of death registration has been mentioned as a source of bias in epidemiological studies, and it is easy to find examples of uncritical use of such data (Carroll 1975). For diseases where there are differences in diagnostic criteria internationally, special studies of disease incidence/prevalence may be undertaken, using similar definitions and criteria in the different participating centers. Examples are international studies of cardiovascular disease and its determinants (the MONICA project, Tunstall-Pedoe et al. 1988) and of asthma (ISAAC study, Pearce et al. 1993).

National boundaries have not always been based on levels of exposure to environmental risk factors, nor of the genetic homogeneity of the populations within them. Thus, study of populations within, and sometimes across, national boundaries has been particularly informative. The geographical units of study can be as small as compatible with a sufficient number of events to generate stable disease rates within them and availability of information on the population-at-risk. Issues of comparability, that may be a source of bias in international comparisons, are usually less important, because recording procedures are in general more similar within countries than between them.

5.5.2.1 Mapping

“Spot maps,” which show the location of individual cases/deaths, have a long history, initially in the investigation of the epidemiology of infectious diseases, more recently as an adjunct to the investigation of clusters of disease (Sect. 5.5.4.1). For non-communicable diseases, the more familiar method is the chloropleth (thematic) map, which uses distinct shading or color to geographical units; usually these are administrative or statistical areas. The reasons for presenting data on risk by place of residence as a map rather than a table are not simply aesthetic. As well as conveying the actual value associated with a particular area, a map conveys a sense of the

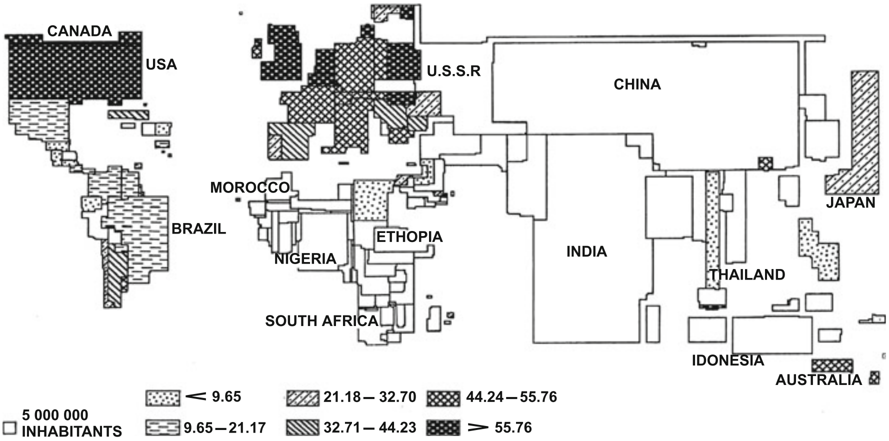


Fig. 5.11 Cartogram of mortality from cancers of the trachea, bronchus, and lung (1981–1984) (Howe 1977)

overall geographical pattern of the mapped variable and allows comparison between the patterns on different maps. This is especially valuable when used to suggest possible causative hypotheses. There are now a large number of disease atlases available, for individual countries (USA, Pickle et al. 1987; UK, Gardner et al. 1984; France, Salem et al. 1999; China, Editorial Committee 1979) or for regions where it is considered that issues of comparability between the participating countries can be overcome (Europe, WHO 1997; Baltic region, Pukkala et al. 2001).

There are a number of specific technical issues to be considered:

- *Choice of Map*

  As well as the various map projections, the *cartogram* has been used by some authors (see Fig. 5.11) (e.g., Verhasselt and Timmermans 1987; Howe 1977). This allocates to the units of study an area proportional to their population size. The idea is to draw attention to the relative numerical importance of the differences displayed, but the resulting maps generally appear somewhat bizarre.
- *Choice of Geographical Unit*

  As noted above, the size of the unit for study is a compromise between the need to provide as much geographical detail as possible (so as to show up any pockets of high or low risk) and achieving stable rates (small variance), so that any spatial patterns are not obscured by random variation.
- *Choice of Parameter*

  The functions plotted are usually rates or ratios. The basic problem is how to compromise between illustrating the actual value of the rate/ratio (generally age standardized in some way) in the different units, which may be influenced by random variation, and giving more weight to those areas with lower variance in the statistic that are unlikely to be due to chance. It is often the case that sparsely populated units cover large geographical areas, while densely populated cities are small. Mapping may well result in impressive high or low

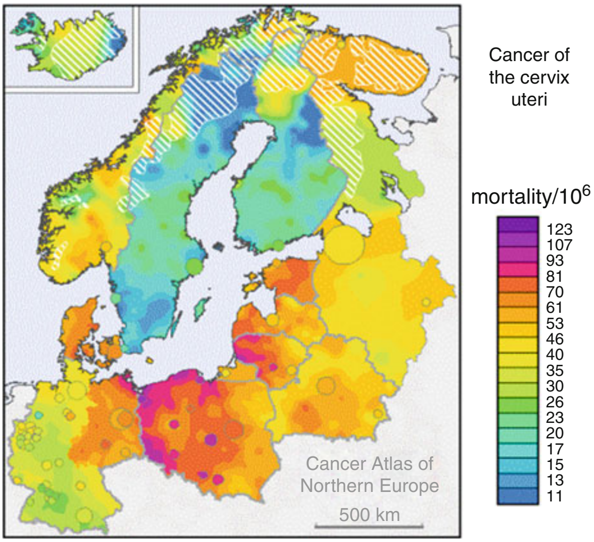


Fig. 5.12 Mortality rates from cancer of the cervix uteri in Northern Europe (Pukkala et al. 2001)

values for eye-catchingly big areas. Plotting the p-values as level of "statistical significance" is not the answer, as this simply highlights differences between the populous units, even when the magnitude of the difference (reflecting its biological significance) is small.

There have been various attempts to circumvent this fundamental problem. The US Atlas of Cancer Mortality used a scale that was a combination of the relative value (e.g., top 10th percentile) and its statistical significance. There is no logical solution to the problem of how to present such maps (is a relative risk of 2 with $p = 0.002$ more or less impressive than an $\hat{R}R$ of 4 and $p = 0.05$?). A different approach is to try to reduce the random variation of rates in small units by assuming that rates in adjacent units will tend to be similar, under the assumption that there is an underlying geographical pattern. Pukkala et al. (2001) color their maps by giving a value to each unit that is the weighted (by distance) average of the other units within 200 kilometers (Fig. 5.12). A more formal method is to plot empirical Bayes estimates of the rates, whereby the values for the units (areas) that are imprecise are improved by estimates from other appropriate areas (Clayton and Kaldor 1987).

- *Choice of Range*

  The number of classes into which to divide the range of values is a compromise between detail and clarity. Usually, five to ten classes are used. There are various choices for the class intervals to be used. The simplest is to use constant intervals (equal steps), in which the range of values is divided into a number of categories of equal size; this works well if the distribution of the data between the units is relatively even but otherwise may be dominated by extreme values, even leaving some classes with no entries at all. If the dataset displays an approximately

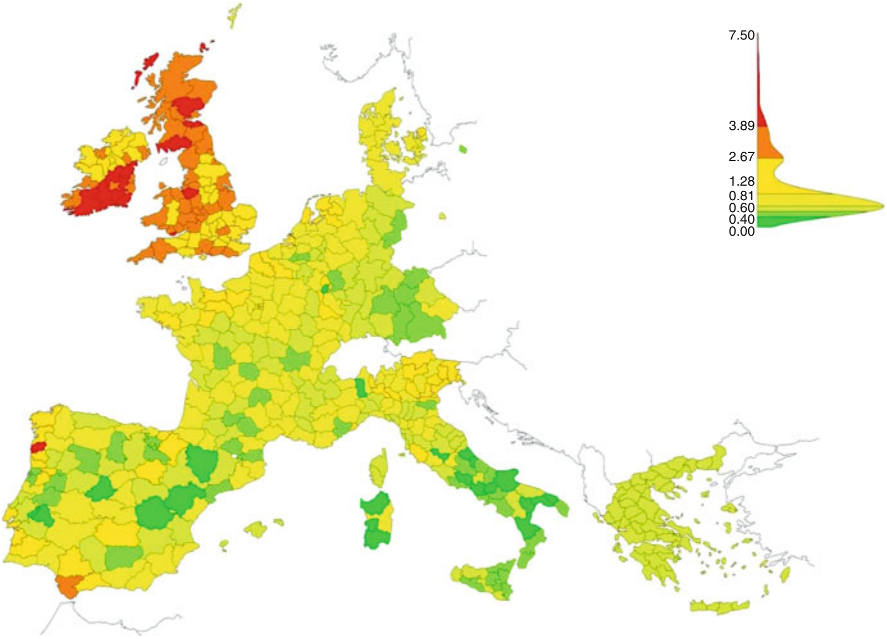


Fig. 5.13 Mortality from esophageal cancer in Europe 1971–1978 (females) (Smans et al. 1992)

normal frequency distribution, class intervals may be standard deviation values; it is useful if the idea of the map is to illustrate deviations from a mean value. Natural divisions of the scale may be used, based upon low points observed in the actual frequency distribution or on some prior knowledge or hypothesis of important dividing values. A relative scale, based on percentiles of the units being mapped, results in irregular variable intervals, but a predictable number of values in each class. The percentiles do not need to be even – in the Scottish Cancer Atlas, the percentiles were 5th, 15th, 35th, 65th, 85th, and 95th, which draw attention to areas of both high and low incidence (Kemp et al. 1985). The same scheme was used for the atlas of cancer mortality in Europe (Smans et al. 1992; Fig. 5.13). This approach means that there is no arbitrary selection of values to plot and that the coloring of all maps tends to be about the same. However, it will obscure outlying (very high or very low values).

- *Choice of Shading/Color*
  The change of shading or color should convey as closely as possible the progression of risk. Color maps are more visually pleasing and can convey more information than those in monochrome. The choice of colors to illustrate the gradations of the scale of the map is not arbitrary and ideally should follow a scale based on the sequence of the spectrum and degree of whiteness (chroma) (Smans and Estève 1992).

5.5.2.2 Urban-Rural Comparisons

There have been many studies in which individuals have been classified into urban or rural dwellers, on the basis of some characteristic of their place of residence, usually, its administrative designation or otherwise as a town/city, or on the basis of population density of the administrative areas (Nasca et al. 1980; Friis and Storm 1993; Barnett et al. 1996). Although distinct differences in the risk of various diseases may be observed, the reasons underlying them are generally obscure. Often, the interest may be in the effects of air pollution on health, given that most air pollution (due to traffic, domestic smoke, or industry) will be more intense in urban areas. However, urban-rural classifications of place of residence is an inefficient way to approach this topic, given the multiplicity of covariates involved.

5.5.2.3 Clustering

The topic is briefly introduced in Sect. 5.5.4, where combinations of person, place, and time are considered.

5.5.3 Time

Investigations of the occurrence of diseases over time are standard tools in epidemiological science and public health surveillance. In the context of investigative epidemiology, temporal studies may generate novel etiological hypotheses or provide confirmatory evidence of existing ones. As well as offering a unique possibility to quantify how risk in populations is changing over time, they offer clues as to the underlying determinants of the observation. Changes in the evolution of incidence rates with time usually imply (in the absence of artifacts) consideration of plausible mechanisms of, and changes in, environmental exposures (time lagged by an approximation of the induction period). Excepting large migrational effects, genetic factors only have a minor impact on time trends of disease (MacMahon and Pugh 1970).

Time trends are also of major importance in measuring the impact of disease control and in studying the effects of primary prevention interventions, screening programs, and the efficacy of treatment regimes. The evaluation of implemented programs (planned or unplanned) may take a “before and after” approach to assess the impact of the intervention on incidence or mortality at the population level. In determining the effectiveness of screening (organized or opportunistic), trends in incidence or mortality, dependent on the specific disease under study, targeted population before and after implementation or comparisons between screening and non-screened groups.

5.5.3.1 Time Trends of Routine Data

The strengths and weaknesses of incidence and mortality data in studying time trends is a subject of much debate. There are complexities in examining trends in either measure, and to avoid erroneous conclusions, it is usually necessary to consider the possibility that artifactual changes over time may have in some way distorted the observed trend.

If the mortality rate is used as a surrogate measure of the risk of developing the disease, a strong assumption of constancy over time in the fatality ratio is required. As survival for many diseases has been improving for several decades, it may be inappropriate to utilize mortality trends as proxies for risk other than for the most fatal diseases. Ideally, mortality rates are best utilized as measures of outcome, rather than occurrence in time trend studies.

Generally a description that utilizes several of these indicators serves to clarify their key properties and aid understanding of the underlying disease processes. There are also a number of temporal datasets on putative or known risk factors collected for particular studies based on, for example, repeated surveys or national surveys, that may be of some utility in elucidating observed trends. Correlation analyses that link such data with trends may clarify particular hypotheses but are limited by their coverage and quality, as well as various potential ecological biases (Morgenstern 1998, Sect. 3.4.2).

5.5.3.2 Describing Secular Trends

Time trend data should be analyzed according to the problem under investigation and the structural characteristics of the data. In the field of health monitoring, the goal might be to quantify the recent secular trend. An estimate of the magnitude and direction (the EAPC – estimated annual percentage change) of the trend over a limited period of time (the last 10 years, say) could be obtained using a simple log-linear model. The EAPC is a useful descriptive measure but should be interpreted with some caution. 95% confidence intervals for the slope should always be given and should help in assessing whether the fitted linear trend may have arisen through chance. If there are elements of curvature in the trend, the EAPC will give incorrect and imprecise estimates of the average unit change. In describing recent trend patterns, the particular choice of time points is often arbitrary, and, in the absence of highly stable rates over time, the EAPC may vary according to the period of time nominated. A preferable description might involve some modeling procedure that could identify sudden changes in the long-term trend and, on that basis, estimate the direction and magnitude of the slope for each epoch of time in which rates are relatively stable.

Methods that quantify trends in segments determined by abrupt linear changes in the trend have been devised by Chu et al. (1999) and by Kim et al. (2000), the latter technique having been implemented in a specially written (and freely available) statistical software package entitled "JoinPoint." The joinpoint regression model essentially searches the temporal data for a few continuous linear phases. The procedure is motivated by the problem of determining the number of joinpoints, that is, breaks in time where abrupt linear changes occur, and an estimate of the EAPC between joinpoints. The minimum and maximum number of joinpoints are user-specified in the software package. To determine up to two joinpoints, for example, a model indicating no change is compared against the model containing two joinpoints. If the null hypothesis of no joinpoints is rejected, then the procedure is applied to test the null hypothesis of one joinpoint against the alternative of two joinpoints. Otherwise, the test for the null hypothesis of no change is considered against the alternative of one joinpoint. Further adaptions of such models have been discussed in Clegg et al. (2009).

5.5.3.3 Age-Period-Cohort Analyses

The first use of the term "cohort" is attributed to Frost in a letter written to a colleague in 1935. The note was published posthumously alongside his landmark paper that discussed some insights that could be attained by visually examining age-specific death rates from tuberculosis according to cohorts, members of a community who share the same birth period, rather than simply in the usual cross-sectional way (Frost 1939). The present day usage of cohort obviously extends well beyond the closed or hypothetical sense of the term – *age-period-cohort analysis* is usually employed to describe temporal studies that include birth cohort analyses, to distinguish it from the generic usage of cohorts in prospective studies (Liddell 1988). In general, such an analysis allows one to examine the influence of each of the three time components and the importance of the particular properties they represent:

- *Birth cohort effects*, on the one hand, may relate to birth itself or may approximate factors related to birth only by exerting influences that are shared in the same group as they age together. An examination of rates according to birth cohort may thus give some insight into the nature and intensity of disease-correlated exposures that may vary across successive generations, and has played a vital role in corroborating evidence from other types of epidemiological study. Temporal patterns in environmental risk factors tend to affect particular generations of individuals in the same way as they age together, and are more likely to exert particular influence on earlier stages of disease development.
- *Period effects*, on the other hand, may act as surrogate measures of events that quickly change incidence or mortality with the same order of magnitude regardless of the age group under study. These effects may be the result of planned interventions that act at later stages of the disease process, for example, novel therapies that improve survival in all age groups. More frequently, they are due to artifactual changes over time, for example, changes in ICD revisions or improvements in diagnostic procedures.
- Age is without doubt a powerful determinant of cancer risk, since it parallels the cumulative exposure to carcinogens over time and the accumulation of the series of mutations necessary for the unregulated cell proliferation that leads to cancer.

A graphical representation of the age-specific rates by period and birth cohort is an essential element of the analysis. Rates versus age though common representations are not always helpful in elucidating the importance of each of the time components – without transformation of the y-axis, the rates can be too closely packed to clearly display the changes, while additionally for cohort trends, they are difficult to interpret, in view of the obvious fact that each rate is only observed for a maximum of a few age groups and at the extremes, only for one age band. More informative perhaps is the depiction of the age-specific rates with period and

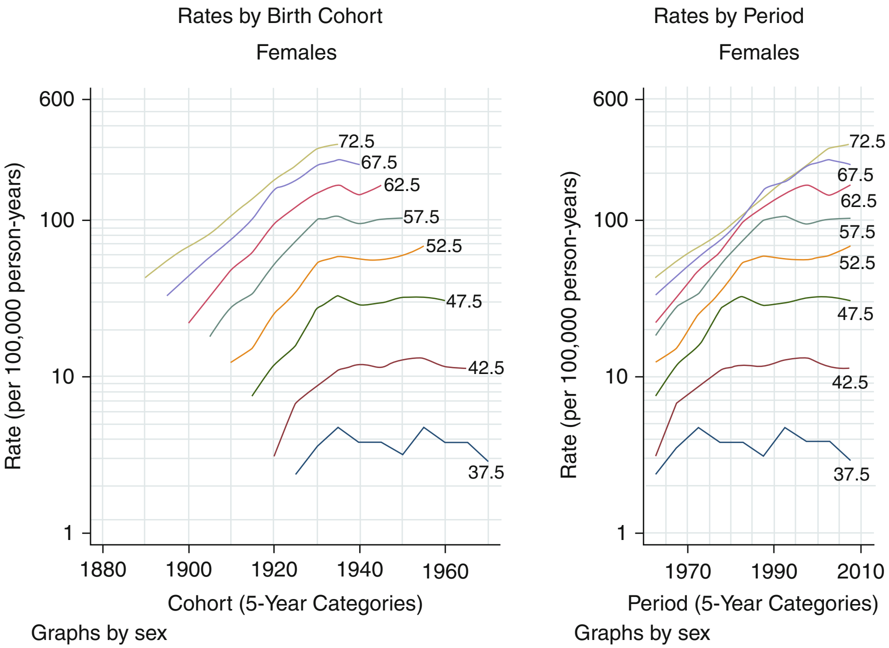


Fig. 5.14 Rates versus calendar period by age group (*left panel*) and rates versus birth cohort by age group, Denmark, lung cancer, females, 1960–2009. The midpoint of the 5-year age groups are indicated in the figures (ANCR (Association of the Nordic Cancer Registries) 2012)

cohort representing the x-coordinates of the two graphs, respectively (Fig. 5.14). Time effects of either origin are therefore apparent when the trends on a semi-log scale are changing in parallel (e.g., by the same relative magnitude) across all age groups in consecutive periods or cohorts.

The effects of the components are sometimes evident to the extent that further investigation may seem unnecessary, given the limited number of variables that require attention. The interpretation of the majority of temporal analyses is however usually more complicated. Rates may fluctuate over time according to the level of random error inherent in the data, dependent on the magnitude of the person-time and the rarity of the disease under investigation. In most situations, the random variation is particularly high in younger persons in recent cohorts. The attribution of changes in the trend to period or cohort effects on the basis of visual means is therefore often not straightforward nor satisfactory, and a comparison of the two-dimensional age-period versus age-cohort graphs can lead to arbitrary opinions as to which component more adequately describes the data. It is in these situations that our understanding of the evolution of cancer risk can be enhanced by the use of more formal statistical procedures. Models offer quantitative and comparable estimates of trend based on objective criteria for choosing the best description of the data, and statistical tests to decide whether the trends are real or random (Estève 1990).

The consequences of subjective judgments based exclusively on graphical descriptions are thus avoided. Statistical models of this nature do not provide definitive answers but offer some guidance as to the importance of each component.

5.5.3.4 The Age-Period-Cohort Model

The emerging importance of birth cohort analyses is in part due to the extensive theoretical and applied research into the age-period-cohort (APC) model in recent times and, importantly, knowledge of its inherent mathematical limitations (Holford 1983; Clayton and Schifflers 1987a, b). The accumulation of available data and advances in statistical theory, including development of the generalized linear model (Nelder and Wedderburn 1972) alongside an increasing availability of software dedicated to fitting such models, also contributes to the development.

As birth cohort is related to a linear function of calendar period and age, the full APC model cannot, given a default set of model constraints, identify all of the parameters of the three components, nor – on the introduction of a further constraint – provide a unique set of estimates. While the APC model is used extensively in applied temporal analyses of disease, the statistical methodologies used to circumvent this problem are numerous and diverse, an indicator of an enduring lack of consensus as to how best to provide satisfactory inferences.

In parallel with advances in statistical theory and computing power, theoretical and applied research on the APC model began to flourish in the late 1970s. During the next couple decades, a number of solutions were offered as how should one present the joint components (e.g., Glenn 1976; Moolgavkar et al. 1979; Day and Charnay 1982; Osmond and Gardner 1982; Holford 1983; Fienberg and Mason 1985; Kupper et al. 1985; Clayton and Schifflers 1987a, b; Tarone and Chu 1992, 1996). A number of reviews and critique of APC models have also been published (e.g., Holford 1998).

The Lexis diagram considers the location of incident cases on one plane, with three time coordinates used to classify events, the date of diagnosis, age at diagnosis, and the date of birth of the individual(s) affected. The third axis, denoted by the diagonal bands crossing the plane from top left to bottom right, represents the date of birth. The APC regression model involves additive contributions of the three time effects on the rate and is given by

$$E[\ln r_{ij}] = \mu + a_i + p_j + c_k$$

where $r_{ij} = Y_{ij} \big/ n_{ij}$ is the incidence (or mortality) rate with n_{ij} as the number of person-time in age group i and period j, assumed fixed and known. a_i is the fixed effect of age group i ($i = 1, 2, \ldots, I$), p_j the fixed effect of period j ($j = 1, 2, \ldots, J$), and c_k the fixed effect of birth cohort with $k = 1, 2, \ldots, K$ where $k = I - i + j$. The number of cancer cases, y_{ij}, is assumed to be distributed as a Poisson random variable with mean λ_{ij}. The model can be estimated readily using maximum likelihood techniques. The numbers of events are fitted via a generalized linear model assuming Poisson errors and a log-link function relating the mean to the linear component. The logarithm of the corresponding person-time is declared as an *offset*, an added constant set to unity for which estimation is not required. The goodness-of-fit is determined as usual by the deviance.

The Identifiability Problem Intrinsic to recognizing the inherent limitations of the APC model is the fact that knowledge of any two factors implies knowledge of the third, making one of the factors redundant. As mentioned above, the index of cohort is defined by the corresponding indexes of age and period, and hence, the three factors are exactly linearly dependent on each other. One further linear constraint must be imposed to ensure the parameter estimates are unique, but the crux of the problem is that this choice of constraint is completely arbitrary in the absence of compelling external information that one can bring to bear in making the selection.

Period and cohort can be considered as weak proxies for our own ignorance regarding the real determinants of time trends (Hobcraft 1985), and it is important that the APC model should be considered as an exploratory tool for investigating the underlying reasons for significant period and cohort effects, adjusted for age. Despite their limitations, such models can render informative results capable of augmenting interpretations based on purely visual approaches.

Classifying Solutions to the Identifiability Problem A number of methods have been proposed that introduce particular constraints on the above APC parameterization so that the identifiability problem appears to be at once resolved, thus yielding unique trend estimates. Such methods often make assumptions founded, necessarily, on mathematical rather than biological principles, that, if inconsistent with reality, induce a bias in all trends, leading to erroneous interpretation. Therefore, such solutions must be carefully scrutinized alongside our present knowledge of the etiology of the cancer under study. Among the methods proposed, a distinction can be drawn that classifies ways of dealing with the analysis and presentation of results from the APC model.

As Holford (1998) points out, a number of quantities can be derived that are estimable and may fulfill the investigative objectives of a temporal study. Such *estimable functions* avoid any imposition of mathematical statements. Rather they are specific reparameterizations that offer summaries of the trends that are identical for any particular set of APC parameters. These conservative but (statistically) correct strategies will be compared with methods that incorporate external data or provide a certain mathematical solution for the cancer outlined, in order to ascertain the level of insight obtained, and the similarities and differences between the methods.

Several authors, notably, Holford (1983) and Clayton and Schifflers (1987b), have noted that certain reparameterizations of the parameters are unique regardless of the constraints imposed, ensuring identifiability, without making any further biological, epidemiological, or mathematical assumptions. Holford (1983) suggested, given the large number of parameters included in the full APC model, for simplicity it is sensible to highlight the non-identifiability in terms of two parameters, one representing a linear function of the three (non-identifiable) slopes and the other, the identifiable curvature of each effect. Clayton and Schifflers (1987a) introduced the term drift or net drift in describing a model for which the two-factor models, age period and age cohort, fit the data equally well.

Drift or δ can be thought of as the average annual change in the rates over time, the passage of time that is common to both axes, calendar period *and* birth cohort, a quantity that cannot be disentangled between the time axes of calendar period and birth cohort. It has become an integral part of the APC modeling strategy, drift being utilized as an estimate of the rate of change of the regular trend, and a partitioner of first-order and curvature effects. The age-drift model implies the same linear change in the logarithm of the rates over time in each age group. Given the linear component over time is identifiable but cannot be allocated in any way to period or cohort, δ can be estimated by either specifying period or cohort as a continuous covariate, and the resulting EAPC estimated as $e^{\delta} - 1$, expressed in the unit of origin.

Perhaps the easiest way to avoid the issue is to ignore the possibility of a three-factor model. However, if such a preference is founded simply on the basis of adequacy of model fit, such an approach may be biased if one of the three effects follows a purely linear pattern (Kupper et al. 1985). In addition, the age-period and age-cohort models are not nested within each other and are therefore not directly comparable. Given its simplicity relative to methods dealing with the three-factor model, however, two-factor models are commonly applied, often when there are a priori beliefs in the nature of the temporal pattern.

5.5.3.5 Predictions of the Future Cancer Burden

A logical extension of the questions "how," "why," and "by how much" posed and potentially answered by the study of temporal analyses is the inquiry as to "what will happen in the future?" The utilization of past trends is a common means of providing such answers and dates back to Andvord (1930), who extrapolated trends in tuberculosis mortality in younger generations into later ages to ascertain the future mortality burden. The response of Frost (1939) – that the exercise was "both tempting and encouraging, but perhaps a little dangerous" – encapsulates the hazardous nature of trend-based disease predictions; we can be reasonably confident that trends observed in the past will not necessarily hold in future.

Still, prediction exercises are highly relevant for administrators requiring accurate information of future numbers of cases to aid health planning (given finite resources), while in public health, future increases or decreases in rates may provide advanced warning of the need for concerted actions or evidence of the enduring success of a specific intervention, planned or otherwise, respectively (Hakulinen 1996).

Predictions of future disease burden are important at the global, regional, and national level. An example of the provision of global predictions of cancer is provided in Fig. 5.15; as the availability of data increases and modeling becomes increasingly possible, predictions should become more accurate (Bray and Møller 2006). The slow decrease in fertility alongside increasing life expectancy projected over the next few decades is having major consequences on the burden of disability and chronic diseases in most populations. On a global scale, taking into account recent trends in different age groups (or birth cohorts) is difficult given the lack

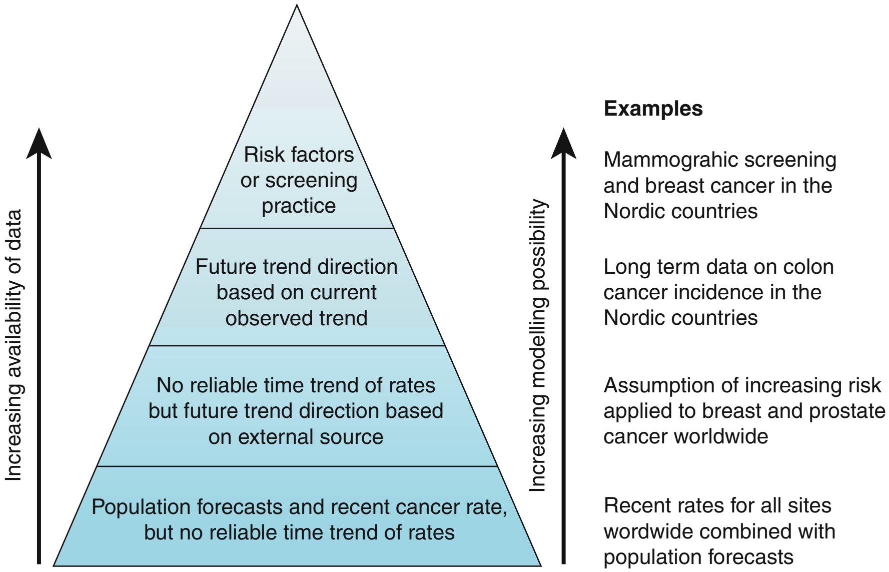

Increasing availability of data
Risk factors or screening practice
Future trend direction based on current observed trend
No reliable time trend of rates but future trend direction based on external source
Population forecasts and recent cancer rate, but no reliable time trend of rates
Increasing modelling possibility
Examples
Mammograhic screening and breast cancer in the Nordic countries
Long term data on colon cancer incidence in the Nordic countries
Assumption of increasing risk applied to breast and prostate cancer worldwide
Recent rates for all sites wordwide combined with population forecasts

Fig. 5.15 Effect of availability of data and flexibility of analytical modeling on cancer prediction strategies (Reproduced from Bray and Møller 2006)

of detailed information on changes in the age-specific incidence in many world areas, and even where such data exist, they are often heterogeneous in different regions. Future scenarios are thus often confined to projecting the effects of probable demographic changes, assuming that future rates remain constant at the present level. As a result of population growth and improved longevity, for example, annual deaths from non-communicable diseases have been projected to rise to 52 million worldwide by 2030 (WHO 2011).

Alongside demographic growth and aging, the changing prevalence and distribution of risk factors (e.g., tobacco smoking, reproductive factors, diet, and unhealthy lifestyles) in different populations makes it important where possible to take into account recent time trends in disease rates so as to improve the accuracy of predictions. If long-term data of reasonable quality are available, APC models (see Sect. 5.5.3) can be used to directly estimate the future burden. Osmond (1985) was one of the first to advocate the projection of the age-, period-, and cohort-specific trends into future time periods to make future predictions of cancer, while Holford (1985) showed that the changes in rates over time, for example, between the projected rate and the rate within the study period (prediction base), were an estimable function, and therefore, predictions did not suffer from a lack of identifiability.

Clayton and Schifflers have however noted that the assumption that drift would remain constant in future periods is a strong assumption that “will rarely be justified in practice,” and several studies have attempted to address this issue by “damping” the projected trend. A Nordic study systematically predicted the number of cases up to 2020 of 20 types of cancer for each sex in the five countries (Møller et al. 2002) based on an APC model that allowed for an increasing reduction in the projection of drift with future time.

Another inherent difficulty is that log-linear models give rise to predictions that grow exponentially over time and, for some cancer types, unrealistically elevated predictions of the number of future events (Hakulinen and Dyba 1994). A power model has been proposed to reduce the growth in the predicted rates and found to improve predictions, as did methods that emphasized trends in the last decade (Møller et al. 2003).

Other methods are available. Berzuini and Clayton (1994) developed a Bayesian APC model to smooth the effects of age, period, and cohort, taking a second difference approach that prevented the rates in adjacent groups from differing too much from each other, as on a multiplicative scale such differences represent adjacent relative risks. Generalized additive models (GAMs) using two-dimensional smoothing splines for age and period have been shown to provide reasonable predictions of female lung cancer rates relative to Bayesian APC models (Clements et al. 2005). Simpler models containing only age and time components have been developed (Hakulinen and Dyba 1994; Dyba et al. 1997; Dyba and Hakulinen 2000) that are particularly useful for data spanning shorter time periods. To avoid the exponential growth over time, models non-linear in parameters but linear in time were applied to cancers with increasing trends (Dyba et al. 1997); these have been shown to perform well compared with other prediction methods (Møller et al. 2002). Log-linear models have also been applied to project new cases of type 1 diabetes in Europe to 2020 (Patterson et al. 2009).

Another approach, the fitting of regression models based on trends in indicators of social and economic development, has been used in projecting cause-specific mortality rates. The methods applied fall into two broad categories (Mathers and Loncar 2006). The most widely used are *aggregate* models which make use of data on historical trends in rates to make future predictions. *Structural* models on the other hand make use of observed correlations between disease rates and independent variables and are the result of projections of these independent variables. This approach has been most widely used in the *Global Burden of Disease* (GBD) exercises, where future mortality rates are predicted based on predictions of the likes of GDP, years of schooling, and prevalence of smoking (Mathers and Loncar 2006).

5.5.4 Combinations

5.5.4.1 Space-Time Clustering

Clustering results from the aggregation of cases (or deaths) in terms of disease group, time, and space, where the number of cases is substantially greater than we would expect when the natural history of the disease and chance fluctuations are taken into account. Investigation of observed clusters, or search for clustering of different diseases, is probably one of the most frequent types of study in descriptive epidemiology. There are, in general, two aims: to identify a possible etiological role for infectious agents in non-communicable diseases and to identify health hazards of sources of environmental contamination, the most popular suspects being sources of pollution (toxic waste dumps, industrial plants) or radiation (nuclear power, electromagnetic fields). Cancer and congenital malformations are the usual subjects of study. It is doubtful if the etiological insights gained are commensurate with the huge volume of research effort (Rothman 1990), although explanations for some clusters of disease have been forthcoming (mesothelioma and sources of asbestos (Baris et al. 1987; Driscoll et al. 1988), mercury and Minamata disease in Japan (Tsuchiya 1992), and dental caries and fluoride in water (Dean 1938)).

Studies of clustering are either a priori or ad hoc. A priori investigation is the search for evidence of clustering in space and/or time in datasets where none is known to exist beforehand but where the investigator may have some prior hypothesis about its existence. Post hoc clusters are observed groupings of events in neighborhoods, schools, occupational units, households, or families, that are considered by someone to be unusual. In both cases, the existence of a point source, responsible for the observed excess, may be suspected.

There are some important considerations in all cluster investigations. The boundaries or units of investigation need to be clearly defined, without reference to the actual observations. The definition should include the disease entity studied, including the diagnostic criteria of a case, the age and sex groups of subjects to be included, and the geographical and temporal boundaries within which the disease event will be counted. Ideally, these are based on biological criteria or on the a priori hypothesis that is being tested. The boundaries should not be defined by the nature of the observations themselves. Demonstrating a clustering effect means that the observed spatial/temporal patterns of disease are significantly different from the expected result, based on an appropriate reference area. Any diligent investigator should be able to achieve this, if sufficient sub-analyses are performed, on varying combinations of diseases, areas, and timescales. Less ideally, post hoc clusters are, by definition, combinations of place, time, and disease that accidentally appeared to be significantly unusual.

A large number of methodological approaches are available for examining datasets for evidence of clustering (see Smith 1982; Marshall 1991; Alexander and Cuzick 1992; Bithell 1992; Alexander and Boyle 1996). Guidelines have been prepared to aid epidemiologists working in public health departments to investigate local post hoc clusters brought to their attention (Centers for Disease Control 1990; Leukaemia Research Fund 1997).

5.5.4.2 Migrant Studies

Migrant studies provide a useful insight into the relative importance of environment and genetic makeup in disease etiology. Disease risk is compared between populations of similar genetic background living in different physical and social environments. Figure 5.16 illustrates the principles; comparison is usually between the rate of disease in migrants (R_{m1}) and the population from which they originated (R_{o}) or between the rates in migrants and those of the new host country in which

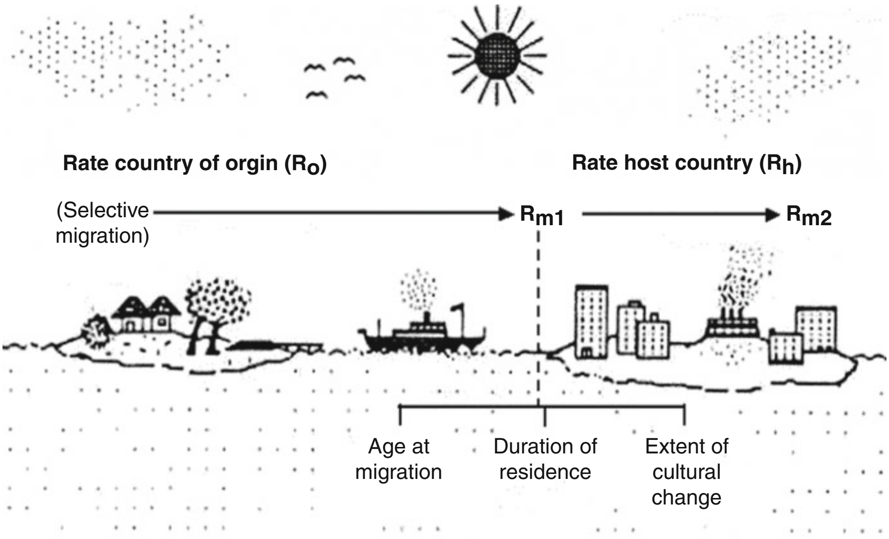

Rate country of orgin (R_o)
Rate host country (R_h)
(Selective migration)
R_{m1}
R_{m2}
Age at migration
Duration of residence
Extent of cultural change

Fig. 5.16 Principles of migrant studies (McMichael and Giles 1988)

they have settled (R_{m1} vs. R_h). The most informative migrant studies are those that permit study of the rate of change of risk following migration. This may be by partitioning migrant rates (R_{m1}) according to age at migration or duration of residence or by comparison of risk in first generation migrants (R_{m1}) and their offspring (R_{m2}).

Studies of disease risk in migrants may involve interview/examination of subjects, in which case information is available not only on the variables of place of origin/place of residence – the focus of migrant studies – but also on covariates. Descriptive studies rely upon data from routine sources (Sect. 5.1.2), and all environmental exposures are subsumed by the variables birthplace and place of residence. Although these may be highly reproducible and subject to little misclassification, in an etiological sense, they are themselves related to numerous more proximate unmeasured "exposures," not only in the external environment (air, soil, water) but also through sociocultural factors (diet, fertility, smoking, etc.), as well as genetic predispositions to them. Adjustment for confounding factors related to birthplace is generally limited, since the variables available are usually few (sex, age, place of residence, occupation, etc.). Nevertheless, as well as the simplicity and convenience that characterize descriptive studies in general, such studies are population based and often of large size (e.g., 120,000 subjects in one study of migrants to Israel; Steinitz et al. 1989). The prospective (cohort) design of most permits risk for several different diseases to be examined.

Migrant studies can only provide useful information when there is a difference in risk between the country of origin (specifically, the population from which the migrants came) and the host population. In the particular instance of the host population comprising the offspring of earlier migrants, there might well be less difference than were it genotypically quite different. Migrants from Spain and Italy might have considerable genetic similarity to the inhabitants of Argentina and Uruguay, for example, countries with populations largely of southern European descent.

Descriptive studies in which “exposure” is investigated in terms of birthplace and residence are not very useful for diseases for which there are obvious causes, with a high population attributable fraction. For example, tobacco smoking is responsible for such a large proportion of cases of lung cancer that risk in migrants will be almost entirely determined by past smoking habits, and the contributions of other environmental factors, for example air pollution, will be quite impossible to evaluate in the absence of detailed knowledge of exposure to tobacco smoke. Conversely, the changes of risk experienced by migrants for cancers of the breast, large bowel, pancreas, and prostate have been far more useful pointers to the relative importance of environmental factors in etiology and to the stage of carcinogenesis at which they may act.

The definition of migrant status is dependent upon the data sources used in a particular study. The most common classification is by place of birth, a relatively well-defined, unchanging attribute, likely to be comparable between the data sources being used (census, vital statistics, registration). Place of birth can also be used in the study of migrants within one country (internal migration). Citizenship or nationality is often recorded on death certificates; it is less useful than place of birth, since migrants will become naturalized to varying degrees, and there are more problems of definition (e.g., dual nationality, stateless persons). Other variables, particularly ethnic group (but also language or religion), have been widely used in comparative studies of populations of different genetic background living in similar environments (or vice versa), in a manner analogous to studies of risk by birthplace (Sect. 5.5.1.3). Studies that employ a combination of ethnic group and birthplace to distinguish first-generation migrants and their offspring are much more informative than either one alone.

The term “environment” embraces, of course, more than the physical surroundings of an individual; it also encompasses all elements of lifestyle that influence disease risk. Thus, while certain aspects of the physical environment (e.g., air and its pollutants, water and trace elements, irradiation-solar, and other forms) change abruptly on migration, other aspects of lifestyle which are related to sociocultural norms will be retained to a greater or lesser degree in the new place of residence. Examples are patterns of diet, childbearing, alcohol and tobacco consumption, sexual habits, and so on. Sociocultural factors also influence the degree of exposure to external environmental agents; thus, given that potential exposure to ultraviolet radiation from the sun is determined by geographical locality, the actual exposures will be modified by culturally defined behavior. As a result, although migrants to countries with sunny climates, such as Israel or Australia, clearly have the same potential for exposure to ultraviolet radiation as the local population, they may be culturally more or less inclined to avoid the sun than the locally born.

The study of disease risk in relation to duration of residence in the new country, or, alternatively, according to the age at the time of migration, is feasible when information is available on the date of migration of the individuals. Provided migrants settle permanently in the host country, age at diagnosis or death is the summation of age at arrival and duration of stay, and it is not possible to evaluate the effect of one of these variables independently of the other. As age is such a strong determinant of risk, and an essential component of any analysis, there is no variability left in duration of stay after controlling for age at arrival, or vice versa; these two variables are therefore inextricably linked. This problem constitutes an extension of the non-identifiability property of age-period-cohort models in the study of time trends (see Sect. 5.5.3.4). A pragmatic solution is to examine each variable in turn (age/duration of residence or age/age at arrival) to see which provides the most plausible pattern of change of risk. "Duration of residence" can be interpreted in terms of dose, that is, assuming that longer periods spent in the new location imply a greater change in cumulative exposure to the relevant etiological factors. It might equally be interpreted in terms of the stage of the disease process at which particular environmental exposures may act. Thus, a rapid change in risk following migration implies change in exposure to a relevant factor and a short period between exposure and disease. Alternatively, the pattern of change may suggest that prolonged exposure is needed before risk is altered or that the agent is only important with respect to exposures early in life. Analysis of risk by "age at migration" may show a clear distinction, in this case, between migrants arriving as children or as adults.

The importance of genetic susceptibility in determining risk is suggested by the persistence of characteristic rates between generations, since the offspring of migrants have been exposed to the environment of the host country for their entire lifespan. However, it is quite likely that they retain some aspects of their parents' lifestyle (as well as their genetic makeup). Some insight into the effect of this can be gained if the rates in the second generation (R_{m2}) can be partitioned according to birthplace of parents (neither, one, or both in the country of origin). These studies require that the data source used contains information on birthplace of parent(s), or ethnicity (if the migrants comprise a distinct ethnic group), or both.

Descriptive studies using routine sources for disease incidence, mortality, or prevalence will have no information on levels of exposure (diet, tobacco, fertility, etc.). However, other data sources may be able to provide population-level data on prevalence or intensity of exposures and permit ecological analyses of risk versus exposure according to birthplace (see Sect. 5.4.2.2). The opportunities for such studies are, unfortunately, limited. Although population surveys may be available, based on interviews (smoking, drinking, dietary habits, reproductive history) or physical examinations (height/weight, blood pressure, blood sugar, etc.), either place of birth is not recorded or, if it is, the samples are usually too small for meaningful results to emerge for any particular group of migrants, who usually comprise only a small fraction of the general population. Occasionally, there have been special ad hoc surveys of migrant populations – and sometimes data from control groups in case-control studies where ethnicity or place of birth has been a major variable of interest (e.g., in studies of diet and cancer in Hawaii (Kolonel et al. 1980; Hankin et al. 1983)). These may provide information on, for example, dietary habits in different migrant groups for comparison with those of the locally born population.

Biases in Migrant Studies

- *Use of Mortality Data*
  Mortality data are normally used as a proxy for incidence (risk of disease); a perfectly valid procedure providing the ratio between mortality and incidence is constant for the groups being compared. This may not be true for international comparisons, since there are known differences in survival between countries (Berrino et al. 2009; Sankaranarayanan et al. 2010). It is less clear whether there are differences in survival by birthplace within a country, although ethnic-specific differences are well documented in the USA (Miller et al. 1996).
- *Data Quality*
  Variation in the quality of data from different sources is particularly troublesome when mortality rates in one country (locally born and migrants) are compared with those from another (country of origin). International variation in completeness and accuracy of death certificate data has been discussed above (Sect. 5.2.4). It may introduce spurious differences in mortality rates. Thus, if the migrant population under study moves from a country with poor certification (of all causes or a specific cause of death) to one with more accurate recording, there will be an apparent increase in the observed rate. Better ascertainment of cause of death, especially for diseases that present diagnostic difficulties, may account for some of the examples of "overshoot" (rates in migrants higher than host country but country of origin rates lower), reported in several studies (Lilienfeld et al. 1972; McMichael et al. 1980).

  Incidence rates from cancer registries are probably more comparable between countries than mortality data. Nevertheless, incidence can be influenced by the detection of asymptomatic cancers during screening, surgery, or autopsy, and is thus related to the extent and nature of such practices. Systematic histological examination of material removed at transurethral prostectomy was responsible for the detection of many "incidental" (non-symptomatic) cancers of the prostate in the USA, and it has been suggested (Shimizu et al. 1991) that the incidence in Japan would have been three to four times higher in the same circumstances. This would explain the apparent rapid increase in the risk of prostate cancer in Japanese migrants to the USA.
- *Mismatching Numerator/Denominator*
  In descriptive studies, person-years at risk are estimated from census data (or from population registers), typically broken down by rather few variables including, in addition to birthplace, age, sex, and place of residence. It is essential that the definition of migrant status is the same in the census and case/death data, but even with the same definition, individuals may be classified in a different way in the two sources. Lilienfeld et al. (1972) present unpublished data on differences between country-of-birth statements on death certificates and census returns for the USA in 1960 – this varied from a 10.8% deficit on death certificates for UK birth to 16.7% excess for Ireland. A source of bias more difficult to detect results from migration that is related to the disease event itself – for example, when migrants return to their country of origin soon before death (so that mortality rates of migrants in the host country are underestimated).

A more practical difficulty in using population-at-risk data results from the fact that censuses are rather infrequent, and interpolations are needed to derive person-years at risk. This can be quite prone to error when several variables are involved, and active migration is still occurring during the study period.

- *Selection Bias*

  Migrant populations are a non-random (self-selected) sample of the population of their country of origin. Very often they come from quite limited geographical areas. For example, migrants to the United States of Italian origin come mainly from the south of that country (Geddes et al. 1993), and a large proportion of US Chinese originate from Guangdong province (King et al. 1985). Alternatively, the migrants may be special social or religious groups with quite distinctive disease patterns. For example, Jews comprised a large proportion of the migrants from Central Europe in the late 1930s and 1940s. Whenever possible, disease rates appropriate to the source population of the migrants should be used for comparisons, rather than the national country of origin rates.

  Migrants are often assumed to be healthier than the average population (the "healthy migrant effect"); this may be because the fact of seeking a new life overseas implies a population that is resourceful and energetic (or at least not chronically ill) or because the sick and disabled are excluded by the immigration authorities of the host country. Conversely, it has been suggested (Steinitz et al. 1989) that permission for Jews to migrate to Israel from countries of the Soviet block was more easily obtained for those in ill health, giving rise to an "unhealthy migrant effect." It is possible to check for the "healthy/unhealthy migrant effect" if risk according to duration of stay in the new country can be estimated. A significant change in rates from those in the host country in recent migrants should suggest this form of bias. Swerdlow (1991) found no sign of any such effect in Vietnamese refugees to England and Wales, and Steinitz et al. (1989) found that exclusion of cancer cases diagnosed within a year of arrival in Israel made no difference to relative risks for short stay (less than 10 years) migrants.

- *Confounding*

  Several demographic variables recorded in the sources of disease information (death certificates, disease registers, etc.) can be considered as confounders – influencing disease risk and associated with exposure (migrant status) – in a study aiming to investigate the effect of birthplace on disease risk. These include date of diagnosis/death, marital status, place of residence, and possibly ethnic group, occupation, and socioeconomic status (such as employment status, income, and educational level).

  Migrants are in the first place rarely distributed homogeneously in their new host country: they tend to settle in certain areas, generally in urban areas, and the establishment of a migrant "colony" in a place tends to attract later migrants to settle there. It may well be inappropriate therefore to compare disease rates in migrants with the entire population of the host country. Table 5.6 illustrates an example of confounding by place of residence. Polish migrants to Argentina live mainly in Buenos Aires (81.2%, compared with 48% of the local born), where mortality rates from colon and breast cancer are higher than elsewhere.

Table 5.6 Confounding by place of residence in a study of cancer mortality in Polish migrants to Argentina (95% confidence intervals in brackets)

Cancer mortality and place of residence	Buenos Aires	Elsewhere in Argentina
Relative risk of colon cancer (M)	1.9	1.0
Relative risk of breast cancer (F)	1.4	1.0
Place of residence and birthplace	Buenos Aires	Elsewhere in Argentina
Born in Poland	81.2%	18.8%
Born in Argentina	47.9%	52.1%
Cancer mortality and birthplace [*Relative risk in Poland-born* vs. *Argentina-born (1.0)*]	Crude	Adjusted for place of residence
Colon cancer (M)	1.34 (1.06–1.68)	1.16 (0.95–1.43)
Breast cancer (F)	0.90 (0.75–1.08)	0.82 (0.63–1.05)

Adjustment for place of residence reduces the relative risk of both cancers, and for colon cancer, the difference from the local born is no longer statistically significant.

Social class and occupation are also known to be strong determinants of disease risk, and it is often clear from census data that migrants are over-represented in specific occupational categories and are atypical of the general population in their socioeconomic profile. Meaningful comparisons should therefore take the social dimension into account.

Temporal trends in incidence or mortality of disease may also be different in the migrant population and in the host country. When data from a long time period are used, the relative risk between them may differ according to time period. This is particularly troublesome when the effect of duration of stay is being studied, since, in general, data from more recent time periods will contain more migrants with long periods of residence than those from earlier years: an adjustment for time period is thus necessary.

Examples of Migrant Studies

- *Single-comparison studies* are the least informative, showing differences in risk between migrants and the locally born but providing no information on the populations from which the migrants came. This may be the consequence of absence of appropriate sources of data (e.g., no accurate mortality statistics) or that rates of disease are unavailable for the appropriate population subgroups from which the migrants came.

  Figure 5.17 (Marmot et al. 1984) shows data on mortality from hypertensive disease (ICD-8 A82) and from coronary heart disease (A83) in men of different migrant groups in England and Wales. There is very large variation in the former (fivefold) and a rather poor correlation of death rates from these two causes between the populations. For most of the countries of origin of the migrants, there are no available data on mortality.

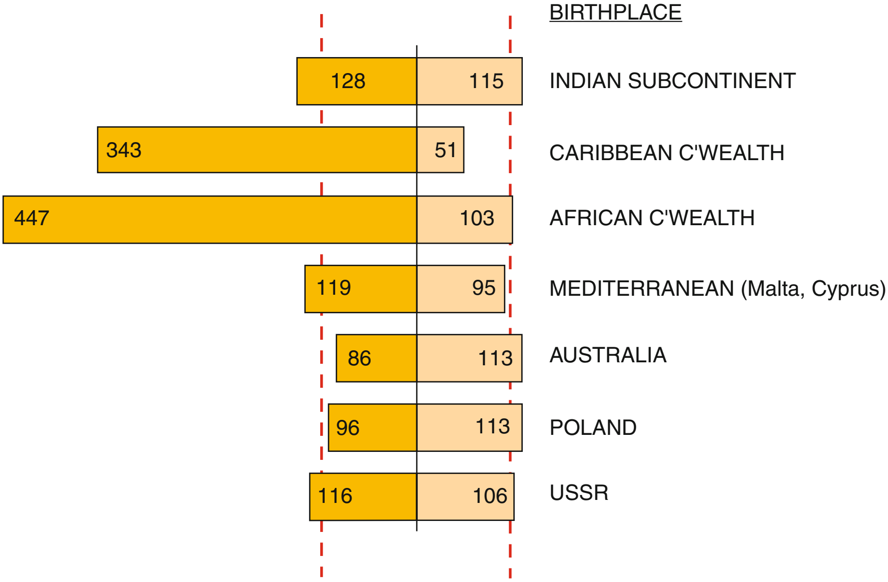


Fig. 5.17 Mortality in migrants to England and Wales, age 20+, 1970–1972 SMR (relative to England and Wales = 100) (Marmot et al. 1984)

- *Two-comparison studies* are the most common type of study reported. They aim to demonstrate the degree to which the risk of a given disease changes in the migrant population away from that in the country of origin and towards that of natives in the new host country. Examples are the several studies that examine mortality rates in populations of predominantly European origin moving to the United States (Haenszel 1961; Lilienfeld et al. 1972), Australia (Armstrong et al. 1983; McCredie 1998), England and Wales (Marmot et al. 1984; Grulich et al. 1992), and South America (Matos et al. 1991; De Stefani et al. 1990). The fact that the data for the two comparisons came from different sources must always be borne in mind, and bias in the first (country of origin) probably explains some of the findings in published studies.

  Figure 5.18 shows results from a study of migrant mortality in the United States (Lilienfeld et al. 1972), comparing age-specific rates in migrants with those in US born whites, and in the countries of origin of the migrants. One of the most impressive differences between migrants and country of origin is for Italians. It is probably in part the result of selection bias: Italian migrants originated mainly from southern Italy, where stomach cancer rates are much lower than in the north (or for the country as a whole) (Geddes et al. 1993).
- *Studies with a time dimension* are studies of the effect of duration of residence or age at migration. Relatively few published studies have been able to study cancer rates in first-generation migrants by duration of stay (or age at arrival) in

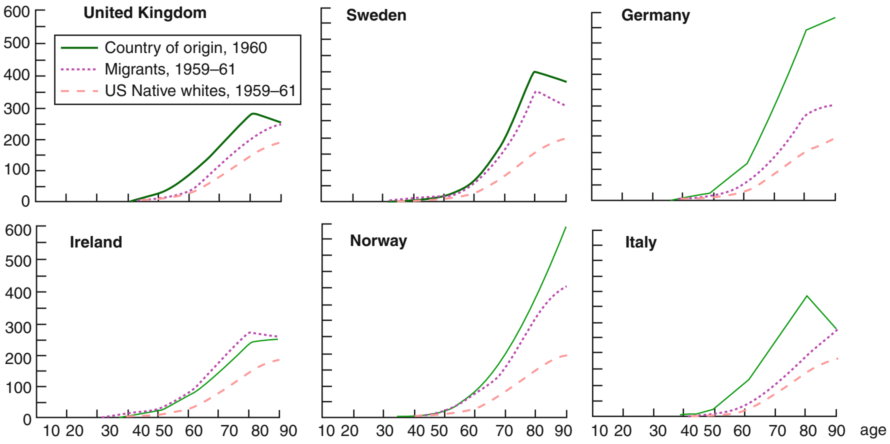


Fig. 5.18 Mortality from cancer of stomach (rate per 10^5) in migrants to United States (Lilienfeld et al. 1972)

the host country. The routine recording in Australian death certificates of date of migration has permitted several studies of mortality in relation to duration of residence in Australia; the findings in relation to gastrointestinal cancers (McMichael et al. 1980) and to malignant melanoma (Khlat et al. 1992) are of particular interest. Figure 5.19 (Khlat et al. 1992) shows the risk of death from melanoma of six migrant populations in relation to either duration of stay or age at arrival, using the Australia-born as the reference group. Since these two variables are completely interdependent (long durations of stay are associated with early ages at arrival), it is impossible to separate their effects. The figure gives the impression that arrival in childhood is associated with relatively high risks but that in age groups 15–24 years, and 25 years and above, risk remains significantly lower than that of the Australia-born. The irregular increase in risk with duration of stay, with a relatively sharp increase after 30 years for many of the groups, makes little biological sense and could well reflect the excess of childhood immigrants in the long stay category.

The Israel Cancer Registry records date of migration for all cases of cancer. This has allowed the risk of cancer to be examined for different populations of migrants in relation to their duration of residence in Israel (Steinitz et al. 1989; Parkin and Iscovich 1997). Figure 5.20 illustrates the risk of cervical cancer in migrants to Israel, relative to the local born, in relation to duration of stay. The data is derived from a long time period (1961–1981) during which there were marked temporal trends in the risk of cervical cancer and in particular a striking increase in incidence (2.5 times) in the Israel-born but little change or even slight declines in risk for the migrant groups. Because most migrations took place before the data collection period, duration of stay is strongly confounded by

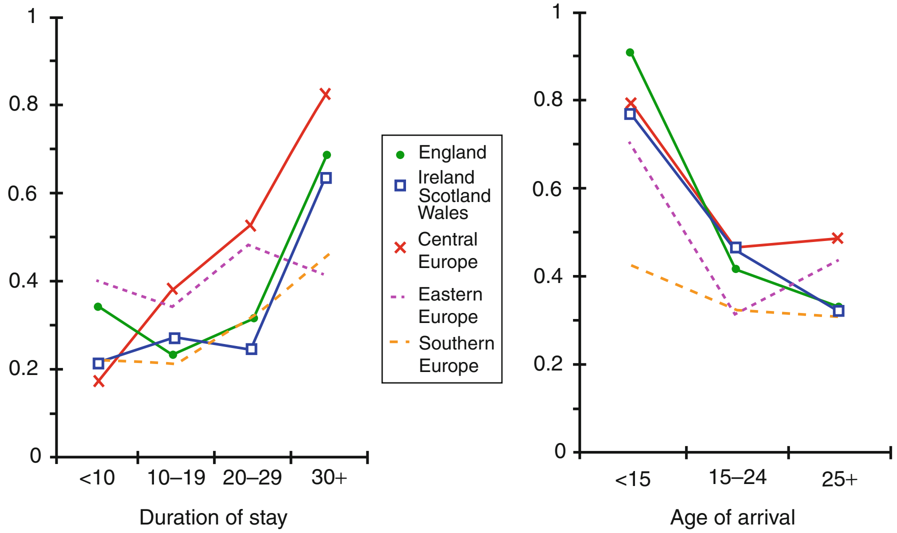


Fig. 5.19 Estimated relative risks of melanoma in male immigrants to Australia, by region of birth and according to duration of stay and age at arrival (both in years), compared with the Australian-born and adjusted for age, period, cohort, and state: Australia, 1964–1985 (Khlat et al. 1992)

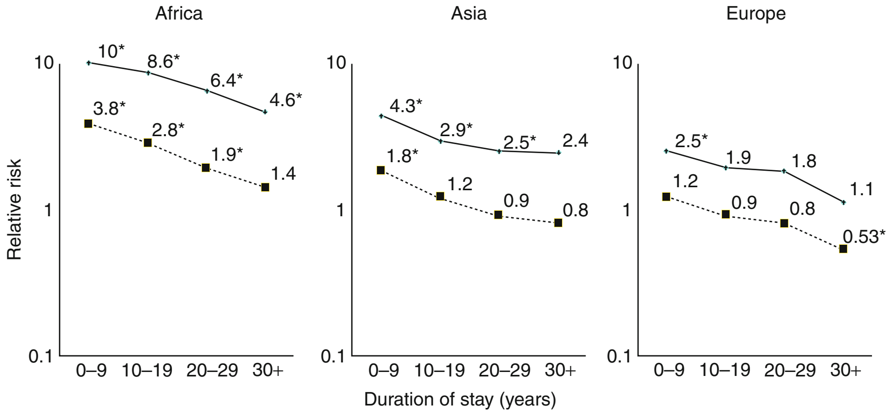
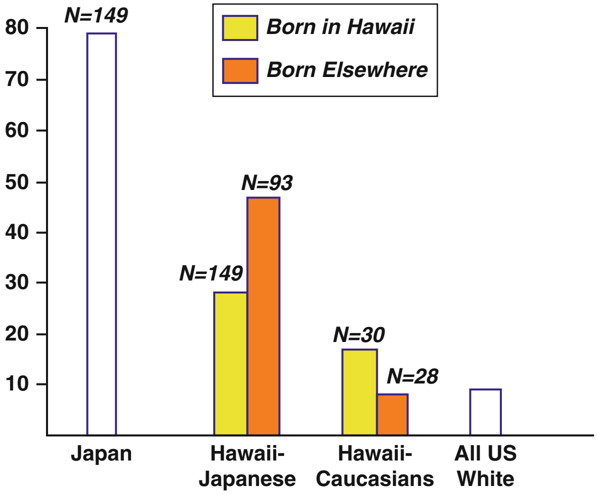


Fig. 5.20 Risk of cervix cancer in migrants (relative to local born) (Parkin et al. 1990)
+—+ Adjusted for age at diagnosis
▪ - - - ▪ Adjusted for age and period of diagnosis

time period (short duration-of-stay cases come mainly from earlier periods, and vice versa), and adjustment for time period has a very striking effect on relative risks (Fig. 5.20). These observations may be explicable in terms of cumulative exposure to Pap smear testing following migration, since the high risk of cervical cancer in these populations is well known to clinicians.

Fig. 5.21 Age-adjusted stomach cancer incidence in Hawaii-Japanese and Caucasians by place of birth, 1973–1977 (Kolonel et al. 1981)



- *Studies of second and subsequent generations of migrants.* The best known studies are those of Japanese in the USA (Haenszel and Kurihara 1968; Locke and King 1980) and Hawaii (Kolonel et al. 1980; Hankin et al. 1983) and of Chinese in the USA (King et al. 1985), distinguishing the foreign-born (first-generation migrants) from the USA-born (their offspring).

  Figure 5.21 shows incidence rates of stomach cancer in two ethnic groups in Hawaii-Japanese and Caucasian (white) in relation to place of birth and provides rates for the populations of the countries of origin, Japan and USA. Incidence rates in Japanese migrants to Hawaii are lower than in Japan, and in Hawaii-born Japanese, they are lower still but still higher than in the white population. Conversely, in the white population of Hawaii, there is an increase in stomach cancer risk in the locally born compared to US whites (or migrants from the USA).

  A number of migrant studies examining cancer risk have already been conducted in Scandinavia. Myrup et al. (2008) linked a cohort of over two million males born after 1930 and living in Denmark 1968–2003, with information in the Danish Civil Registration System including immigration histories to investigate testicular cancer risk. First- and second-generation immigrants had relative risks of testicular cancer of 0.37 and 0.88, respectively, compared with men born in Denmark of parents born in Denmark. The rate in first-generation immigrants was not modified by age at immigration or duration of stay, reflecting rather their country of origin. Together these findings implicate that early-in-life exposures are a major determinant of later risk of this disease.

  Hemminki and Li used the nationwide Swedish Family-Cancer Database to examine cancer risks related to migration based on over 600,000 adult immigrants to Sweden. Comparing cancer rates between immigrants and the locally born in Sweden (Hemminki et al. 2002), all cancer rates were 5 and 8% lower among immigrant men and women, respectively. However, male immigrants had a 41% increased risk of lung cancer. Using the same database to examine second-generation migrants (Hemminki and Li 2002) based on descendants of an immigrant father or mother, all cancer rates were marginally lower among the offspring of migrants relative to the host population. Comparison of risk between the first- and the second-generation migrants suggests that the first 2 decades of life are influential in determining the pattern of cancer development in later life.

Birthplace of parents, which is sometimes recorded on death certificates or cancer registries, has been little used to study cancer risk in offspring of migrants. Balzi et al. (1995) used mortality data from Canada to study cancer risks in Italian migrants and Canadian-born individuals of Italian parentage. The latter group was separated into either those with two parents born in Italy or only one. Figure 5.22 shows results for the two most common cancers, stomach and lung. The risk of stomach cancer, which in migrants is two to three times that in the reference population (Canada-born of Canadian parents), is no longer significantly raised in their offspring, while trends in the opposite direction are seen for lung cancer. Parkin and Iscovich (1997), using data from the Israel Cancer Registry, presented odds ratios for migrants and the Israel-born population according to parents' birthplace. Individuals with parents from North Africa retained the increased risk of nasopharynx cancer seen in migrants from that area, while the risk of melanoma remained low in migrants from this area and their offspring.

- *Studies that include information on exposures* at population-level data for the migrants and the population of the host country, and sometimes for country of origin, are examples of ecological studies (Sect. 5.4.2.2).

  McMichael and colleagues (1980) related mortality rates from gastrointestinal cancers in European migrants to Australia with per capita food consumption data from a period 10 years earlier in Australia and the countries of origin; results from a national dietary survey were included later (McMichael and Giles 1988).

  For the Japanese population of Hawaii, the dietary intake of a range of nutrients can be estimated from the control subjects in case-control studies and from special surveys (Kolonel et al. 1980; Hankin et al. 1983), and the Japanese population (as well as Hawaii whites) can be separated into those born in Hawaii or elsewhere. The Japan-Hawaii heart study also provides a large amount of information on dietary patterns in Japan and Hawaii (Kagan et al. 1974). These data have been used to help in the interpretation of the changes in risk of stomach cancer in Japanese migrant populations (Fig. 5.21). Thus, second-generation Japanese eat less pickled vegetables and dried salted fish than Japanese migrants (born in Japan), whereas the whites who were born in Hawaii seem to eat these foods more frequently than whites born elsewhere. Both of these items have been associated with an increase in the risk of stomach cancer in case-control studies. The observations were similar for consumption of rice (and total carbohydrate intake), also suggested as etiologically important in some studies.

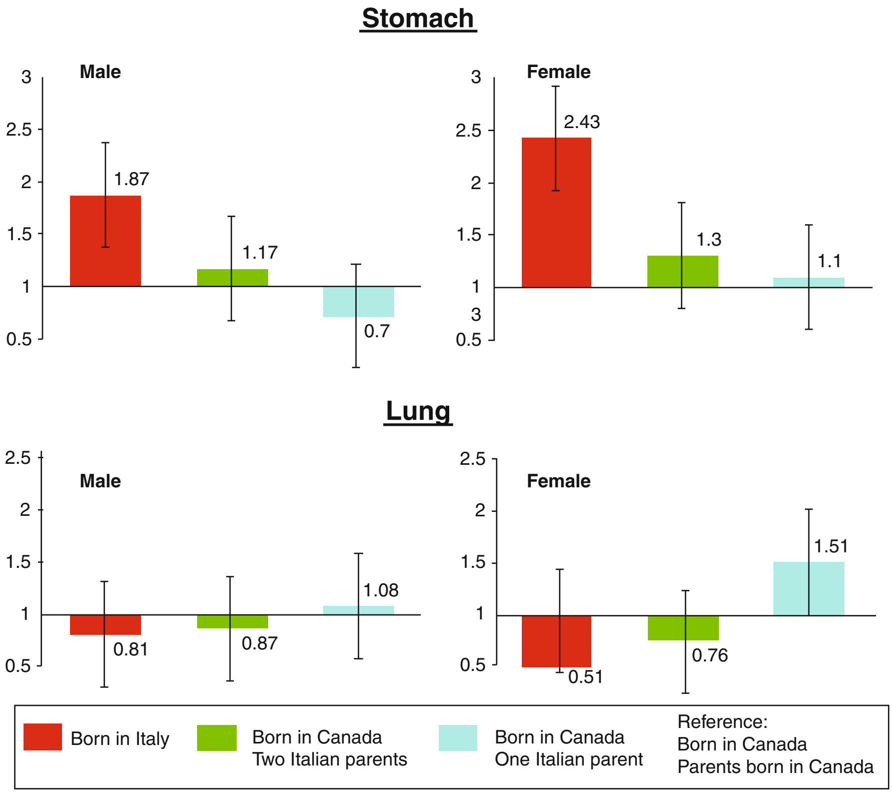


Fig. 5.22 Risks and 95% confidence intervals of stomach and lung cancer in Italian migrants and Canada-born individuals of Italian parentage (Reference population: Canada-born of Canadian parents, Balzi et al. 1995)

5.6 Conclusions

Studies of disease patterns using registers of vital events are often cited as the foundations of modern epidemiology – the work of Graunt on "Bills of Mortality" (Graunt 1662) foreshadowing the enormous contribution of Farr, as statistician to the Registrar General, in analyzing the material on cause of death, provided by the routine registration of vital events in England and Wales. Descriptive epidemiology is a continuation of this theme. The increasing availability of databases related to the health of individuals, or to their possible exposure to causes of disease, has enormously increased the scope for investigations providing clues to etiological associations. As we describe in this chapter, the information on individuals that is contained in personalized databases is rarely closely related to pathogenetic mechanisms, so that observed associations will generally be suggestive only of causative pathways, and a stimulus to more focused investigations. Thus, differences in the risk of disease according to locality (place of residence, or of birth, or both) may be quite suggestive of an underlying cause (e.g., in contaminants of water, background radiation, soil mineral deficiencies), but the hypotheses will require testing, if possible, by studies that involve collection of information from individual subjects, on variables that are related to the hypothesized more proximal causes.

This chapter provides some background on the tools available in descriptive epidemiology: The sources of routine data on health status, and the basics of measurement and comparison. It illustrates how these can be applied in the study designs familiar to epidemiologists, and provides a series of examples, illustrating the principles of descriptive studies.

In an era of genomics, proteomics, metabolomics, etc., the study of disease risk according to demographic characteristics of individuals, or their place of birth, or residence, and the evolution of risk over time may seem mundane, with little relevance to unraveling the secrets of life. However, the relative simplicity of descriptive studies, in which large populations can be investigated, at relatively little expense, means that they will continue to be widely used. The ecological study, for all its defects, is probably the most popular study design in epidemiology (and much beloved of other disciplines too, e.g., in economics). Even within the framework of etiological research, some of the most basic observations remain a challenge to biological explanation (e.g., the striking sex ratios for some cancers).

Descriptive epidemiology has, of course, a wide application beyond the realm of the academic epidemiologist focused on investigating the causation of disease. Epidemiology, is, after all, concerned with "the application of this study [of the distribution and determinants of health-related states or events in specified populations] to control of health problems" (Sect. 5.1.1). Planning and evaluation of health care requires a knowledge of the magnitude of different problems – their distribution in subgroups within the community, their past and likely future evolution, and their amenability to different interventions – and monitoring of the effectiveness of interventions, be they in prevention, early diagnosis, or therapy. Epidemiology is the keystone of public health, and attention to the well-being of the health of the populace requires a sound knowledge of the principles of descriptive epidemiology, rather than a detailed knowledge of the proteomics of gene interactions. It remains the basic toolkit of the community-orientated health specialist.

References

Alderson M (1981) International mortality statistics. McMillan, London

Alexander FE, Boyle P (eds) (1996) Methods for investigating localised clustering of disease. IARC scientific publications no. 135. International Agency for Research on Cancer, Lyon

Alexander FE, Cuzick J (1992) Methods for the assessment of disease clusters. In: Elliot P, Cuzick J, English D, Stern R (eds) Geographical and environmental epidemiology: methods for small-area studies. Oxford University Press, Oxford, pp 238–250

ANCR (Association of the Nordic Cancer Registries) (2012) NORDCAN on the web. http://ancr.nu. Accessed 17 Sept 2012

Andersen A, Barlow L, Engeland A, Kjaerheim K, Lynge E, Pukkala E (1999) Work-related cancer in the Nordic countries. Scand J Work Environ Health 25(Suppl 2):1–116

Andvord KF (1930) What can we learn from studying tuberculosis by generations? Norsk Mag Loegevidensk 91:642–660

Armstrong BK, Woodings TL, Stenhouse NS, McCall MG (1983) Mortality from cancer in migrants to Australia-1962 to 1971. University of Western Australia, Nedlands, p 138

Armstrong BK, White E, Saracci R (1994) Principles of exposure measurement in epidemiology. Oxford University Press, Oxford

Balzi D, Geddes M, Brancker A, Parkin DM (1995) Cancer mortality in Italian migrants and their offspring in Canada. Cancer Causes Control 6:68–74

Baquet CR, Ringen K (1986) Cancer among blacks and other minorities: statistical profiles. National institute of health publication 86–2785. National Cancer Institute/USDHHS, Washington, DC

Baris I, Simonato L, Artvinli M, Pooley F, Saracci R, Skidmore J, Wagner C (1987) Epidemiological and environmental evidence of the health effects of exposure to erionite fibres: a four-year study in the Cappadocian region of Turkey. Int J Cancer 39:10–17

Barnett E, Strogatz D, Armstrong D, Wing S (1996) Urbanisation and coronary heart disease mortality among African Americans in the US South. J Epidemiol Community Health 50: 252–257

Beral V, Chilvers C, Fraser P (1979) On the estimation of relative risk from vital statistical data. J Epidemiol Community Health 33:159–162

Berrino F, Verdecchia A, Lutz JM, Lombardo C, Micheli A, Capocaccia R (2009) EUROCARE Working Group. Comparative cancer survival information in Europe. Eur J Cancer 45:901–908

Berzuini C, Clayton D (1994) Bayesian analysis of survival on multiple time scale. Stat Med 13:823–838

Biggar RJ, Horm J, Goedert JJ, Melbye M (1987) Cancer in a group at risk of acquired immunodeficiency syndrome (AIDS) through 1984. Am J Epidemiol 126:578–586

Bithell J (1992) Statistical methods for analysing point source exposures. In: Elliot P, Cuzick J, English D, Stern R (eds) Geographical and environmental epidemiology: methods for small-area studies. Oxford University Press, Oxford, pp 221–230

Bray F, Møller B (2006) Predicting the future burden of cancer. Nat Rev Cancer 6(1):63–74

Bray F, Guilloux A, Sankila R, Parkin DM (2002) Practical implications of imposing a new World standard population. Cancer Causes Control 13:175–182

Bray F, Ren JS, Masuyer E, Ferlay J (2013) Global estimates of cancer prevalence for 27 sites in the adult population in 2008. Int J Cancer 132(5):1133–1145

Brenner H, Savitz DA, Jöckel KH, Greenland S (1992) Effects of nondifferential exposure misclassification in ecologic studies. Am J Epidemiol 135:85–95

Breslow NE (1984) Elementary methods of cohort analysis. Int J Epidemiol 13:112–115

Breslow NE, Day NE (1975) Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. J Chronic Dis 28:289–303

Breslow NE, Day NE (1980) Statistical methods in cancer research, vol I: the analysis of case-control studies. IARC scientific publications no. 32. International Agency for Research on Cancer, Lyon

Breslow NE, Day NE (1987) Statistical methods in cancer research, vol II: the design and analysis of case-control studies. IARC scientific publications no. 82. International Agency for Research on Cancer, Lyon

Carroll KK (1975) Experimental evidence of dietary factors and hormone-dependent cancers. Cancer Res 35:3374–3383

Carstairs V (1995) Deprivation indices: their interpretation and use in relation to health. J Epidemiol Community Health 49(Suppl 2):S3–S8

CDC-Centers for Disease Control and Prevention (2012) Surveys and data collection systems. http://www.cdc.gov/nchs/surveys.htm. Accessed 14 Sept 2012

Centers for Disease Control (1990) Guidelines for investigating clusters of health events. Morbidity & Mortality Weekly Report 39(No. RR-11):1–23

Chu KC, Baker SG, Tarone RE (1999) A method for identifying abrupt changes in U.S. cancer mortality trends. Cancer 86:157–169

Clayton D, Kaldor J (1987) Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. Biometrics 43:671–681

Clayton D, Schifflers E (1987a) Models for temporal variation in cancer rates. I: age-period and age-cohort models. Stat Med 6:449–467

Clayton D, Schifflers E (1987b) Models for temporal variation in cancer rates. II: age-period-cohort models. Stat Med 6:469–481

Clegg LX, Hankey BF, Tiwari R, Feuer EJ, Edwards BK (2009) Estimating average annual per cent change in trend analysis. Stat Med 28(29):3670–3682

Clements MS, Armstrong BK, Moolgavkar SH (2005) Lung cancer rate predictions using generalized additive models. Biostatistics 6(4):576–589

Cochran WG (1954) Some methods for strengthening the common χ^2 tests. Biometrics 10: 417–451

Curado MP, Edwards B, Shin HR, Ferlay J, Heanue M, Boyle P, Storm H (2007) Cancer incidence five continents, vol IX. IARC scientific publications No. 160. International Agency for Research on Cancer, Lyon

Curtin LR, Mohadjer LK, Dohrmann SM, Montaquila JM, Kruszan-Moran D, Mirel LB, Carroll MD, Hirsch R, Schober S, Johnson CL (2012) The National Health and Nutrition Examination Survey: sample design, 1999–2006. Vital Health Stat 2(155):1–39

Cutler SJ, Ederer F (1958) Maximum utilization of the life table method in analyzing survival. J Chronic Dis 8:699–712

Das Gupta P, Sen A, Marglin S (1972) Guidelines for project evaluation. United Nations, New York

Day NE (1992) Cumulative rate and cumulative risk. In: Parkin DM, Muir CS, Whelan SL, Gao Y-T, Ferlay J, Powell J (eds) Cancer incidence in five continents, vol VI. IARC scientific publications no. 120. International Agency for Research on Cancer, Lyon

Day NE, Charnay B (1982) Time trends, cohort effects and aging as influence on cancer incidence. In: Magnus K (ed) Trends in cancer incidence – causes and practical implications. Hemisphere Publication Corporation, Washington, DC, pp 51–65

De Stefani E, Parkin DM, Khlat M, Vassalo A, Abella M (1990) Cancer in migrants to Uruguay. Int J Cancer 46:233–237

Dean HT (1938) Endemic fluorosis and its relation to dental caries. Public Health Rep 53: 1443–1452

Dempsey M (1947) Decline in tuberculosis: the death rate fails to tell the entire story. Am Rev Tuberc 86:157–164

Doll R, Peto R (1981) The causes of cancer. Oxford University Press, Oxford

Doll R, Payne P, Waterhouse J (eds) (1966) Cancer incidence in five continents: a technical report. Springer, New York

Driscoll RJ, Mulligan WJ, Schultz D, Candelaria A (1988) Malignant mesothelioma. A cluster in a Native American pueblo. New Engl J Med 318:1437–1438

Dwyer T, Hetzel BS (1980) A comparison of trends of coronary heart disease mortality in Australia, USA and England & Wales with reference to three major risk factors – hypertension, cigarette smoking and diet. Int J Epidemiol 9:67–71

Dyba T, Hakulinen T (2000) Comparison of different approaches to incidence prediction based on simple interpolation techniques. Stat Med 19(13):1741–1752

Dyba T, Hakulinen T, Paivarinta L (1997) A simple non-linear model in incidence prediction. Stat Med 16:2297–2309

Ederer F, Axtell LM, Cutler SJ (1961) The relative survival rate: a statistical methodology. Natl Cancer Inst Monogr 6:101–121

Editorial Committee for the Atlas of Cancer Mortality (1979) Atlas of cancer mortality in the People's Republic of China. China Map Press, Shanghai

Estève J (1990) International study of time trends. Some methodological considerations. Ann N Y Acad Sci USA 609:77–84

Estève J, Benhamou E, Raymond L (1994) Statistical methods in cancer research, vol IV. Descriptive epidemiology. IARC scientific publications no. 128. International Agency for Research on Cancer, Lyon

Fienberg SE, Mason WM (1985) Specification and implementation of age, period, and cohort models. In: Mason WM, Fienberg SE (eds) Cohort analysis in social research: beyond the identification problem. Springer, New York, pp 44–88

Friis S, Storm H (1993) Urban-rural variation in cancer incidence in Denmark 1943–1987. Eur J Cancer 29A:538–544

Frost WH (1939) The age selection of mortality from tuberculosis in successive decades. Am J Hyg 30:91–96

Gardner MJ, Winter PD, Taylor CP, Acheson ED (1984) Atlas of cancer mortality in England & Wales, 1968–1978. Wiley, Chichester

Geddes M, Parkin DM, Khlat M, Balzi D, Buiatti E (eds) (1993) Cancer in Italian migrant populations. IARC scientific publication no. 123. International Agency for Research on Cancer, Lyon

Gillum RF (2002) New considerations in analyzing stroke and heart disease mortality trends: the year 2000 age standard and the international statistical classification of diseases and related health problems, 10th revision. Stroke 33:1717–1721

Glenn NK (1976) Cohort analysts' futile quest: statistical attempts to separate age, period and cohort effects. Am Sociol Rev 41:900–904

Graunt J (1662) Natural and political observations mentioned in a following index, and made upon the bills of mortality. Roycroft, London

Greenland S, Morgenstern H (1989) Ecological bias, confounding and effect modification. Int J Epidemiol 18:269–274

Grulich AE, Swerdlow AJ, Head J, Marmot MG (1992) Cancer mortality in African and Caribbean migrants to England and Wales. Br J Cancer 66:905–911

Haenszel W (1961) Cancer mortality among the foreign born in the United States. J Natl Cancer Inst 26:37–132

Haenszel W, Kurihara M (1968) Studies of Japanese migrants. I. Mortality from cancer and other diseases among Japanese in the United States. J Natl Cancer Inst 40:43–68

Hakulinen T (1996) The future cancer burden as a study subject. Acta Oncol 35(6):665–670

Hakulinen T, Dyba T (1994) Precision of incidence predictions based on Poisson distributed observations. Stat Med 13:13–23

Hankin JR, Kolonel LN, Yano K, Heilbrun L, Nomura AMY (1983) Epidemiology of diet related diseases in the Japanese migrant population of Hawaii. Proc Nutr Soc Aust 8:22–40

Heasman MA, Lipworth L (1966) Accuracy of certification of cause of death. Studies in medical and population subjects, vol 20. Her Majesty's Stationery Office, London

Hemminki K, Li X (2002) Cancer risks in second-generation immigrants to Sweden. Int J Cancer 99(2):229–237

Hemminki K, Li X, Czene K (2002) Cancer risks in first-generation immigrants to Sweden. Int J Cancer 99(2):218–228

Hirsch A (1883) Handbook of geographical and historical pathology, vols I–III (trans: Second German Edition by Creighton C). The New Sydenham Society, London

Hoar SK, Morrison AS, Cole P, Silverman DT (1980) An occupation and exposure linkage system for the study of occupational carcinogenesis. J Occup Med 22:722–726

Hobcraft J (1985) Age, period, and cohort effects in demography: a review. In: Mason WM, Fienberg S (eds) Cohort analysis in social research: beyond the identification problem. Springer, New York, pp 89–135

Holford TR (1983) The estimation of age, period and cohort effects for vital rates. Biometrics 39:311–324

Holford TR (1985) An alternative approach to statistical age-period-cohort analysis. J Chronic Dis 38:831–840

Holford TR (1998) Age-period cohort analysis. In: Armitage P, Colton T (eds) Encyclopaedia of biostatistics. Wiley, Chichester, pp 82–99

Howe GM (1977) A world geography of human diseases. Academic, London/New York/San Francisco

Howe GR, Lindsay JP (1983) A follow-up study of a ten-percent sample of the Canadian labor force. I. Cancer mortality in males, 1965–73. J Natl Cancer Inst 70:37–44

Kagan A, Harris BR, Winkelstein W, Johnson KG, Kato H, Syne SL, Rhoads GG, Gay ML, Nichaman MZ, Hamilton HB, Tillotson J (1974) Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: demographic, physical, dietary and biochemical characteristics. J Chronic Dis 27:345–364

Kaldor J, Khlat M, Parkin DM, Shiboski S, Steinitz R (1990) Log linear models for cancer risk among migrants. Int J Epidemiol 19:233–239

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481

Kemp I, Boyle P, Smans M, Muir C (eds) (1985) Atlas of cancer in Scotland 1975–1980. IARC scientific publications no. 72. International Agency for Research on Cancer, Lyon

Khlat M, Vail A, Parkin DM, Green A (1992) Mortality from melanoma in migrants to Australia: variation by age at arrival and duration of stay. Am J Epidemiol 135:1103–1113

Kim HJ, Fay MP, Feuer EJ, Midthune DN (2000) Permutation tests for joinpoint regression with applications to cancer rates. Stat Med 19:335–351

King H, Li JY, Locke FB, Pollack ES, Tu JJ (1985) Patterns of site-specific displacement in cancer mortality among migrants: the Chinese in the United States. Am J Public Health 75:237–242

Kirkwood BR, Sterne JAC (2003) Essential medical statistics, 2nd edn. Blackwell, Oxford

Kogevinas M, Pearce N, Susser M, Boffetta P (eds) (1997) Social inequalities and cancer. IARC scientific publications no. 138. International Agency for Research on Cancer, Lyon

Kolonel LN, Nomura AM, Hirohata T, Hankin JH, Hinds MW (1981) Association of diet and place of birth with stomach cancer incidence in Hawaii Japanese and Caucasians. Am J Clin Nutr 34:2478–2485

Kolonel LN, Hinds MW, Hankin JR (1980) Cancer patterns among migrant and native-born Japanese in Hawaii in relation to smoking, drinking and dietary habits. In: Gelboin LJV, MacMahon B, Matsushima T, Sugimura T, Takayama S, Takebe H (eds) Genetic and environmental factors in experimental and human cancer. Japan Scientific Societies Press, Tokyo, pp 327–340

Koopman JS, Longini IM Jr (1994) The ecological effects of individual exposures and non-linear disease dynamics in populations. Am J Public Health 84:836–842

Korteweg R (1951) The age curve of lung cancer. Br J Cancer 5:21–27

Kuh DL, Ben-Shlomo Y (1997) A life course approach to chronic disease epidemiology; tracing the origins of ill-health from early to adult life. Oxford University Press, Oxford

Kupper LL, McMichael AJ, Symon MJ, Most BM (1978) On the utility of proportional mortality analysis. J Chronic Dis 31:15–22

Kupper LL, Janis JM, Karmous A, Greenberg BG (1985) Statistical age-period-cohort analysis: a review and critique. J Chronic Dis 38:811–830

Last JM (ed) (2001) A dictionary of epidemiology, 4th edn. Oxford University Press, New York

Layard R, Glaister S (1994) Cost-benefit analysis, 2nd edn. Cambridge University Press, Cambridge

Leukaemia Research Fund Centre for Clinical Epidemiology (1997) Handbook and guide to the investigation of clusters of diseases. University of Leeds, Leeds

Lexis W (1875) Einleitung in die Theorie der Bevölkerungsstatistik. Karl J. Trübner, Strassburg

Liddell FD (1988) The development of cohort studies in epidemiology: a review. J Clin Epidemiol 41:1217–1237

Lilienfeld AM, Lilienfeld DE (1980) Foundations of epidemiology. Oxford University Press, New York

Lilienfeld AM, Levin ML, Kessler II (1972) Mortality among the foreign born and in their countries of origin. In: Lee HP, Lilienfeld AM, Levin ML, Kessler II (eds) Cancers in the United States. Harvard University Press, Cambridge, pp 233–278

Locke FB, King H (1980) Cancer mortality among Japanese in the United States. J Natl Cancer Inst 65:1149–1156

Logan WPD (1982) Cancer mortality by occupation and social class, 1851–1971. IARC scientific publications no. 36 & studies on medical and population subjects no. 44. International Agency for Research on Cancer/Her Majesty's Stationery Office, Lyon/London

Lynge E, Thygesen L (1990) Occupational cancer in Denmark. Cancer incidence in the 1970 census population. Scand J Work Environ Health 16(Suppl 2):3–35

Macbeth H, Shetty P (eds) (2001) Health and ethnicity. Taylor & Francis, London/New York

MacMahon B, Pugh H (1970) Epidemiology: principles and methods. Little Brown and Company, Boston

MacMahon B, Trichopoulos D (1996) Epidemiology: principles and methods, 2nd edn. Little Brown and Company, Boston

Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 22:719–748

Marmot M (1999) Epidemiology of socioeconomic status and health: are determinants within countries the same as between countries? Ann N Y Acad Sci 896:16–29

Marmot MG, Adelstein AM, Bulusu L (1984) Immigrant mortality in England and Wales 1970–1978: causes of death by country of birth. Studies on medical and population subjects no. 47. Her Majesty's Stationery Office, London, p 144

Marshall RJ (1991) A review of methods for the statistical analysis of spatial patterns of disease. J R Stat Soc A 154:S421–S441

Mathers CD, Loncar D (2006) Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med 3(11):e442

Mathers CD, Fat DM, Inoue M, Rao C, Lopez AD (2005) Counting the dead and what they died from: an assessment of the global status of cause of death data. Bull World Health Organ 83:171–177

Matos E, Khlat M, Loria DI, Vilensky M, Parkin DM (1991) Cancer in migrants to Argentina. Int J Cancer 49:805–811

McCredie M (1998) Cancer epidemiology in migrant populations. Recent results. Cancer Res 154:298–305

McGill HC (ed) (1968) Geographic pathology of atherosclerosis. Williams & Williams, Baltimore

McMichael AJ, Giles GG (1988) Cancer in migrants to Australia: extending the descriptive epidemiological data. Cancer Res 48:751–756

McMichael AJ, McCall MG, Hartshorne JM, Woodings TL (1980) Patterns of gastrointestinal cancer in European migrants to Australia: the role of dietary change. Int J Cancer 25:431–437

Miettinen OS, Wang JD (1981) An alternative to the proportionate mortality ratio. Am J Epidemiol 114:144–148

Miller BA, Kolonel LN, Bernstein L, Young JL, Swanson GM, West D, Key CR, Liff JM, Glover CS, Alexander GA (eds) (1996) Racial/ethnic patterns of cancer in the United States 1988–1992. NIH publication no. 96–4104. National Cancer Institute, Bethesda

Moger TA, Haugen M, Bray F, Grotmol T, Tretli S, Aalen OO (2009) Frailty modeling of bimodal age-incidence curves of nasopharyngeal carcinoma in low-risk populations. Biostatistics 10(3):501–514

Møller B, Fekjær H, Hakulinen T, Tryggvadóttir L, Storm HH, Talbäck M, Haldorsen T (2002) Prediction of cancer incidence in the Nordic countries up to the year 2020. Eur J Cancer Prev Suppl 11(1):S1–S96

Møller B, Fekjær H, Hakulinen T, Sigvaldason H, Storm HH, Talbäck M, Haldorsen T (2003) Prediction of cancer incidence in the Nordic countries: empirical comparison of different approaches. Stat Med 22(17):2751–2766

Moolgavkar SH, Stevens RG, Lee JA (1979) Effect of age on incidence of breast cancer in females. J Natl Cancer Inst 62:493–501

Morgenstern H (1982) Uses of ecologic analysis in epidemiologic research. Am J Public Health 72:1336–1344

Morgenstern H (1998) Ecologic studies. In: Rothman KJ, Greenland S (eds) Modern epidemiology, 2nd edn. Lippincott-Raven, Philadelphia

Morrow RH, Bryant JH (1995) Health policy approaches to measuring and valuing human life: conceptual and ethical issues. Am J Public Health 85:1356–1360

Murray CJL, Lopez AD (eds) (1996) The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020. Harvard University Press, Cambridge

Murray CJL (1994) Quantifying the burden of disease: the technical basis for disability-adjusted life years. Bull World Health Org 72:429–445

Murray JL, Axtell LM (1974) Editorial: impact of cancer: years of life lost due to cancer mortality. J Natl Cancer Inst 52:3–7

Myrup C, Westergaard T, Schnack T, Oudin A, Ritz C, Wohlfahrt J, Melbye M (2008) Testicular cancer risk in first- and second-generation immigrants to Denmark. J Natl Cancer Inst 100:41–47

Nasca PC, Burnett WS, Greenwald P, Brennan K, Wolfgang P, Carlton K (1980) Population density as an indicator of urban-rural differences in cancer incidence, upstate New York, 1968–1972. Am J Epidemiol 112:362–375

Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc A 135:370–384

Office for National Statistics, UK (2012) General Lifestyle Survey overview. A report on the 2010 General Lifestyle Survey (glfreport2010_tcm77-259420.pdf). http://www.ons.gov.uk/ons/search/index.html?newquery=General%20Household%20Survey%20-%20GB. Accessed 07 Aug 2012

Osmond C (1985) Using age, period and cohort models to estimate future mortality rates. Int J Epidemiol 14(1):124–129

Osmond C, Gardner MJ (1982) Age, period and cohort models applied to cancer mortality rates. Stat Med 1:245–259

Parker SL, Davis KJ, Wingo PA, Ries LA, Heath CW Jr (1998) Cancer statistics by race and ethnicity. CA Cancer J Clin 48:31–48

Parkin DM (2006) The evolution of the population-based cancer registry. Nat Rev Cancer 6(8):603–612

Parkin DM, Iscovich JA (1997) The risk of cancer in migrants and their descendants in Israel: II Carcinomas and germ cell tumours. Int J Cancer 70:654–660

Parkin DM, Steinitz R, Khlat M, Kaldor J, Katz L, Young J (1990) Cancer in Jewish migrants to Israel. Int J Cancer 45:614–621

Parkin DM, Kramarova E, Draper GJ, Masuyer E, Michaelis J, Neglia J, Qureshi S, Stiller CA (1998) International incidence of childhood cancer, vol II. IARC scientific publications no. 144. International Agency for Research on Cancer, Lyon

Parkin DM, Whelan SL, Ferlay J, Teppo L, Thomas DB (eds) (2002) Cancer incidence in five continents, vol VIII. IARC scientific publications no. 155. International Agency for Research on Cancer, Lyon

Patterson CC, Dahlquist GG, Gyürüs E, Green A, Soltész G, EURODIAB Study Group (2009) Incidence trends for childhood type 1 diabetes in Europe during 1989–2003 and predicted new cases 2005–20: a multicentre prospective registration study. Lancet 373(9680):2027–2033

Pearce N, Weiland S, Keil U, Langridge P, Anderson HR, Strachan D, Bauman A, Young L, Gluyas P, Ruffin D, Crane J, Beasley R (1993) Self-reported prevalence of asthma symptoms in children in Australia, England, Germany and New Zealand: an international comparison using the ISAAC protocol. Eur Respir J 6:1455–1461

Percy C, Stanek E, Gloeckner L (1981) Accuracy of cancer death certificates and its effect on cancer mortality statistics. Am J Public Health 71:242–250

Peto R, Doll R (1997) There is no such thing as aging. Brit Med J 315:1030–1032

Piantadosi S, Byar DP, Green SB (1988) The ecological fallacy. Am J Epidemiol 127:893–904

Pickle LM, Mason TJ, Howard N, Hoover R, Fraumeni JF Jr (1987) Atlas of US cancer mortality among whites, 1950–1980. DHHS publications No. (NIH) 87–2900. US Government Printing Office, Washington, DC

Pocock SJ, Cook DG, Beresford SAA (1981) Regression of area mortality rates on explanatory variables; what weighting is appropriate? Appl Stat 30:286–295

Porta M (ed) (2008) A dictionary of epidemiology, 5th edn. IEA Oxford University Press, Oxford

Prentice RL, Sheppard L (1990) Dietary fat and cancer; consistency of the epidemiological data, and disease prevention that may follow from a practical reduction in dietary fat consumption. Cancer Causes Control 1:81–97

Puffer RR, Wynne-Griffith G (1967) Patterns of urban mortality, vol 151. Pan American Health Organization Scientific Publication, Washington, DC

Pukkala E, Söderman B, Okeanov A, Storm H, Rahu M, Hakulinen T, Becker N, Stabenow R, Bjarnadottir K, Stengrevics A, Gurevicius R, Glattre E, Zatonski W, Men T, Barlow L (2001) Cancer atlas of northern Europe. Publication no. 62. Cancer Society of Finland, Helsinki

Registrar General's Office for England and Wales (1978) Occupational mortality: The Registrar General's Decennial Supplement for England and Wales, 1970–72, XVIII. Series D S 1. HMSO No. [O 11 69064483]. Her Majesty's Stationery Office, London

Rose G (1982) Incubation period of coronary heart disease. Br Med J 284:1600–1601

Rose G, Marmot M (1981) Social class and coronary heart disease. Br Heart J 45:13–19

Rosen G (1993) A history of public health. Johns Hopkins University Press, Baltimore

Rothman KJ (1990) A sobering start for the cluster busters conference. Am J Epidemiol 132:6–11

Rothman KJ, Greenland S (eds) (1998) Modern epidemiology, 2nd edn. Lippincott-Raven, Philadelphia

Salem G, Rican S, Jougla E (1999) Atlas de la Sante en France. John Libbey Eurotext, Paris

Sankaranarayanan R, Swaminathan R, Brenner H, Chen K, Chia KS, Chen JG, Law SC, Ahn YO, Xiang YB, Yeole BB, Shin HR, Shanta V, Woo ZH, Martin N, Sumitsawan Y, Sriplung H, Barboza AO, Eser S, Nene BM, Suwanrungruang K, Jayalekshmi P, Dikshit R, Wabinga H, Esteban DB, Laudico A, Bhurgri Y, Bah E, Al-Hamdan N (2010) Cancer survival in Africa, Asia, and Central America: a population-based study. Lancet Oncol 11:165–173

Segi M (1960) Cancer mortality for selected sites in 24 countries (1950–57). Tohoku University of Medicine, Sendai

Seow A, Straughan PT, Ng EH, Emmanuel SC, Tan CH, Lee HP (1997) Factors determining acceptability of mammography in an Asian population: a study among women in Singapore. Cancer Causes Control 8:771–779

Shimizu H, Ross RK, Bernstein L, Yatoni R, Henderson BE, Mack TM (1991) Cancers of the prostate and breast among Japanese and white immigrants in Los Angeles County. Br J Cancer 63:963–966

Siemiatycki J (1996) Exposure assessment in community-based studies of occupational cancer. Occup Hyg 3:41–58

Siemiatycki J, Day NE, Fabry J, Cooper JA (1981) Discovering carcinogens in the occupational environment: a novel epidemiologic approach. J Natl Cancer Inst 66:217–225

Smans M, Estève J (1992) Practical approaches to disease mapping. In: Elliott P, Cuzick J, English D, Stern R (eds) Geographical and environmental epidemiology: methods for small-area studies. Oxford University Press, Oxford, pp 141–150

Smans M, Muir CS, Boyle P (1992) Atlas of cancer mortality in the European Economic Community. IARC scientific publications no. 107. International Agency for Research on Cancer, Lyon

Smith GD, Leon D, Shipley MJ, Rose G (1991) Socioeconomic differentials in cancer among men. Int J Epidemiol 20:339–344

Smith P (1987) Comparison between registries: age-standardized rates. In: Muir C, Waterhouse J, Mack T, Powell J, Whelan S (eds) Cancer incidence in five continents, vol V. IARC scientific publications no. 88. International Agency for Research on Cancer, Lyon

Smith PG (1982) Spatial and temporal clustering. In: Schottenfeld D, Fraumeni JF (eds) Cancer epidemiology and prevention. WB Saunders Comp, Philadelphia. Chapter 21

Soerjomataram I, Lortet-Tieulent J, Ferlay J, Forman D, Mathers C, Parkin DM, Bray F (2012a) Estimating and validating disability-adjusted life years at the global level: a methodological framework for cancer. BMC Med Res Methodol 12(1):125

Soerjomataram I, Lortet-Tieulent J, Ferlay J, Mathers C, Parkin DM, Forman D, Bray F (2012b) Disability-adjusted life years: country-specific estimates for 27 cancers in 12 world regions. Lancet 380(9856):1840–1850

Steenland K, Beaumont J (1984) The accuracy of occupation and industry data on death certificates. J Occup Med 26:288–296

Steinitz R, Parkin DM, Young JL, Bieber CA, Katz L (1989) Cancer incidence in Jewish migrants to Israel 1961–1981. IARC scientific publications no. 98. International Agency for Research on Cancer, Lyon

Swerdlow A (1991) Mortality and cancer incidence in Vietnamese refugees in England and Wales: a follow-up study. Int J Epidemiol 20:13–19

Tarone RE, Chu KC (1992) Implications of birth cohort patterns in interpreting trends in breast cancer rates. J Natl Cancer Inst 84:1402–1410

Tarone RE, Chu KC (1996) Evaluation of birth cohort patterns in population disease rates. Am J Epidemiol 143:85–91

Tsuchiya K (1992) The discovery of the causal agent of Minamata disease. Am J Ind Med 21: 275–280

Tunstall-Pedoe H for WHO MONICA Project Principal Investigators (1988) The World Health Organisation MONICA Project (Monitoring Trends and Determinants in Cardiovascular Disease): a major international collaboration. J Clin Epidemiol 41:105–114

United Nations (1984) Handbook of household surveys. United Nations, New York

Verhasselt Y, Timmermans A (1987) World maps of cancer mortality. Geografisch Instituut VUB, Brussels

Walter SD (1991) The ecologic method in the study of environmental health. II Methodologic issues and feasibility. Environ Health Perspect 94:67–73

WHO (1992) International statistical classification of diseases and related health problems, 10th revision. World Health Organization, Geneva

WHO (1997) Atlas of mortality in Europe: subnational patterns 1980/1981 and 1990/1991. WHO regional publications; European series no. 75, World Health Organization, Geneva

WHO (1998) Reported information on mortality statistics. World health statistics annual 1996. World Health Organization, Geneva, pp xvi–xxi

WHO (2000) The World Health Report 2000 Health Systems: improving performance, World Health Organization, Geneva

WHO (2008) The global burden of disease: 2004 update. http://www.who.int/healthinfo/global_burden_disease/GBD2004_DisabilityWeights.pdf. Accessed 13 Dec 2012

WHO (2011) Global status report on noncommunicable diseases 2010. http://www.who.int/nmh/publications/ncd_report_full_en.pdf. Accessed 13 Dec 2012

WHO (2012) Global Health Observatory (GHO): World health statistics 2012. http://www.who.int/gho/publications/world_health_statistics/2012/en/. Accessed 14 Sept 2012

Wigle DT, Mao Y, Howe G, Lindsay J (1982) Comparison of occupation on survey and death records in Canada. Can J Public Health 73:242–247

Cohort Studies

6

Anthony B. Miller, David C. Goff Jr., Karin Bammann, and Pascal Wild

Contents

A.B. Miller (✉)
Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

D.C. Goff Jr.
Department of Epidemiology, Colorado School of Public Health, Aurora, CO, USA

K. Bammann
Institute for Public Health and Nursing Research, Faculty of Human and Health Sciences,
University of Bremen, Bremen, Germany

P. Wild
Direction scientifique (Scientific Management), Occupational Health and Safety Institute (INRS),
Vandoeuvre-lès-Nancy, France

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_6,

6.1 Introduction

This chapter summarizes the basic features of cohort studies, a type of observational epidemiology study that some have also called longitudinal, or prospective, though these terms also apply to other epidemiological designs. A cohort study evaluates both the risk and the rate of disease or disease-related outcomes in a population that is characterized in terms of relevant risk factors or exposures, placed under observation, and followed for some time until disease develops or not. In contrast to its classical counterpart, the case-control study (cf. chapter ▸Case-Control Studies of this handbook), cohort studies can relate multiple diseases to the exposure or exposures identified. On the other hand, cohort studies are frequently restricted to a limited number of exposures and potential confounders that can be included in the study, especially if historical data are used.

The chapter is organized as follows: First, a brief historical perspective on cohort studies is given, showing the importance of this study design by giving examples. Second, conceptual features of cohort studies are presented, where the two basic types of cohort studies, concurrent and non-concurrent historical cohort studies, are summarized, and the basic concepts of data analysis in cohort studies are described. These concepts include the description of outcome events in the cohort, the comparison with external data, and the analysis of effects of exposure. The chapter then deals with key concerns of cohort studies, including selection of the study population and on the important question of how to determine exposure and outcome events in the framework of a cohort study. Ethical issues, mainly raised through the potential future use of specimens, are also discussed. Examples are given on many issues. Although some of these may seem simplistic, they are presented in this way to ensure that general principles are clearly depicted.

6.2 A Brief Historical Perspective on Cohort Studies

Cohort studies have been used for over a century to study determinants of disease. Since the early days of epidemiology, they have been used as a powerful tool to study a broad range of exposures such as infections, nutritional factors, occupational factors, and lifestyle factors, as the following examples illustrate.

The classical study on the London cholera epidemic of 1849 conducted by John Snow is an example of a cohort study on infectious diseases (Snow 1855; Sutherland 2002). Reports from the Registrar General had drawn attention to the possibility that differences in water supply were associated with differences in cholera rates across sections of London. Two different water companies (the Lambeth and the Southwark and Vauxhall) supplied households within various regions of London, and frequently these two water companies supplied adjacent households. The companies differed in one important feature, the location of the water intake. The Lambeth had moved their water intake upstream from the sewage discharge point in 1849, whereas the Southwark and Vauxhall continued to obtain water downstream of the sewage discharge point. Dr. Snow classified households according to their exposure to the two water sources and showed a substantial difference in cholera mortality, 37 versus 315 cholera deaths per 10,000 households served by the Lambeth and Southwark and Vauxhall companies, respectively.

Cohort studies continue to be an important tool in the investigation of infectious diseases. For example, McCray (1986) used a cohort design to quantify the risk of developing the acquired immunodeficiency disorder (AIDS) among healthcare workers exposed to blood and body fluids of AIDS patients.

Joseph Goldberger employed a variety of epidemiological approaches, including cohort methods, to study pellagra, a systemic disease endemic in the southeast of the United States in the late nineteenth and early twentieth century (Terris 1964). In one investigation, Goldberger examined the dietary exposures of households in relation to the occurrence of pellagra and demonstrated that a cornmeal subsistence diet was associated with pellagra. Subsequent trials showed that pellagra could not be transmitted from person to person, as might be expected for an infectious disease, but could be prevented by the “pellagra-preventive factor,” later determined to be niacin. More recently, Oomen et al. (2001) studied the association of trans-fatty acids, a hydrogenation product of oils containing polyunsaturated fatty acids, and heart disease among men in the Netherlands. They found a relative risk of 1.28 of heart disease for an increase of 2% of energy from trans-fatty acid intake at baseline.

Occupational epidemiology is another classical field of application of cohort studies. Typically, workers exposed to a putative harmful substance are compared to other workers in the industry or to the general population. Occupational cohorts were used to study, for example, the association between exposure to dyes and urinary bladder cancer (Case et al. 1954), exposure to mustard gas and respiratory cancer (Wada et al. 1968), and exposure to benzene and leukemia (Rinsky et al. 1987). The health effects for workers exposed to asbestos have been the subject of many studies (e.g., Liddell et al. 1997; Selikoff et al. 1965) and continue to be examined. Ulvestad and colleagues (2004) conducted a cohort study of members of the Norwegian Trade Union of Insulation Workers hired between 1930 and 1975 and followed through 2002, demonstrating relative increases in risk of mesothelioma and lung cancer when compared with the experience of the general population.

Lifestyle exposures have also attracted the attention of epidemiologists, including physical activity, tobacco and alcohol use, diet, and nutrition. Morris and colleagues (1953a, b) demonstrated that British bus drivers had approximately twice the risk of heart disease in comparison to the more active conductors (who went up and down the stairs to collect tickets). This result was confirmed in a comparison of postmen with telephonists and clerks (Morris et al. 1953a, b). In 1951, Doll and Hill (1954) initiated a cohort study of British physicians by collecting data on tobacco use via questionnaire. By collecting death certificate data, they were able to demonstrate a tenfold increased risk of lung cancer death for smokers compared to non-smokers (Doll and Peto 1976). Doll et al. (1994a) also reported on the association of alcohol consumption with mortality among British doctors demonstrating a u-shaped relationship, with greater mortality among abstainers and heavy drinkers and the lowest mortality among moderate drinkers, defined as

1–2 drinks per day on average. Concerns were raised that the increased risk described in abstainers may be falsely elevated by the experience of former drinkers who may have quit drinking due to health decline. This concern was addressed by Eigenbrodt et al. (2001) using cohort methodology within the Atherosclerosis Risk in Communities (ARIC) study. Eigenbrodt et al. measured perceived health status and alcohol consumption behavior longitudinally and were able to identify changes in health status that preceded changes in drinking behavior. They demonstrated that perceived health decline predicted cessation of drinking, thereby providing evidence that the risk among abstainers may have been inflated in studies that failed to distinguish between lifelong abstainers and former drinkers.

Despite disadvantages regarding cost and complexity, cohort studies remain of substantial public health importance as indicated by several of the previously cited examples and by such evidence as was recently provided by the National Institutes of Health (NIH). The NIH is considering the establishment of a 500,000-person cohort study to examine genetic and environmental influences on common diseases in the United States (National Institutes of Health 2004). Similar large cohort studies have been or are being initiated in Canada, Germany, and the UK. The large sample sizes under consideration for these studies will enable the examination of gene-gene and gene-environment interaction in the general population and in subgroups of interest. Therefore, a sound understanding of cohort methodology is of substantial importance to the modern epidemiologist.

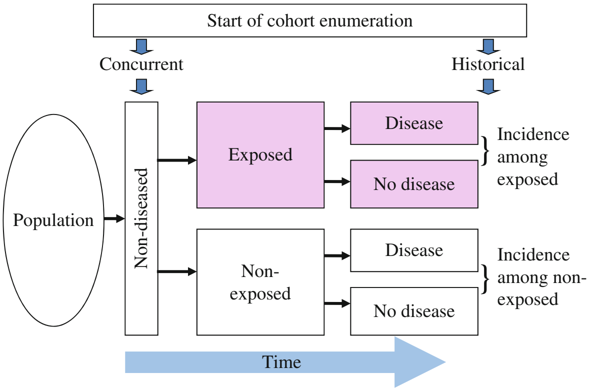
6.3 Conceptual Foundations

6.3.1 Types of Cohort Studies: Concurrent and Non-Concurrent Approaches

The central feature of a cohort study is the collection of exposure data in a defined population and the subsequent surveillance of possible outcome events regarding health, morbidity, and mortality. For this purpose, healthy members of a defined population (the cohort) are classified according to their exposure status (e.g., exposed vs. unexposed) and followed over a long period with respect to their health status. Then, the question can be answered if incidence of outcome events is associated with former presence or absence of exposure, which would indicate a possible causal relationship.

Within this framework, cohort studies can be classified in two major categories depending on the timing of follow-up period relative to the time of study conduct. In *concurrent or prospective cohort studies* (a simplified scheme is given in Fig. 6.1), a defined population is assembled and possibly screened to eliminate persons with disease where the resulting non-diseased population is the so-called population at risk. Then, information on exposure, possible confounders, and other important factors is gathered. The cohort members are subsequently followed into the future recording outcome events of interest. In *non-concurrent* or *historical cohort studies* (Fig. 6.1), a population is assembled from available historical data records, for

Fig. 6.1 Design of a cohort study



example, from company files. Exclusion of persons with preexisting disease defines the population at risk; assessment of exposure and other factors is based on the available data from the past. Cohort members are monitored for outcome events through existing documents and data systems (e.g., vital statistics files or disease registries) from some point in the past to the present. As in concurrent studies, outcome rates may be compared across exposure categories within the cohort, or, if all members of the cohort are assumed to be exposed, outcome rates may be compared between the cohort and the general population, assumed to be the unexposed reference population. A combined approach is also possible, with the cohort assembled and followed initially through historical documents or other data sources such as data from registries and subsequently followed using concurrent methods. Kleinbaum et al. (1982) introduced this as ambispective study design. The distinction between these two major categories of cohort studies has important implications regarding data collection.

In concurrent studies, the methods for cohort assembly and data collection can more easily be controlled, whereas in non-concurrent studies, the investigators must rely on data recorded in historical records, almost always for reasons other than medical research. This notable disadvantage of the non-concurrent approach is compensated by the ability to study exposures, such as occupational exposures, that meet one or more of the following key conditions: (1) the exposure can be attributed to selected employed populations based on individual records of job descriptions or other employment data, (2) the exposure is relatively rare in the general population outside the occupations of interest, (3) the induction period is long, and (4) the health concern is substantial; thus, an early answer can be obtained to questions of risk, an important advantage from a public health perspective. In practice, the design and analysis of cohort studies require advanced methods in order to account for exposures that are not fixed, not dichotomous, and not time invariant. Individuals may move from being exposed to being unexposed and vice versa over time. Advanced models also need to account for dynamic cohorts where subjects enter or leave a population at risk at various points in time during the observation period.

Because many diseases tend to be multifactorial in causation, a crucial point in the validity of cohort studies is the inclusion of data on possible confounders at baseline. This is a problem in historical cohort studies that will be discussed in the section on determining exposures below.

Two modern extensions of cohort studies that try to integrate the advantages of cohort and case-control studies are designed to have nearly all the power of classic cohort studies but utilize relatively economically detailed exposure information from questionnaires or biomarkers or other biological measurements determined from the collection of biological specimens at the time the study is initiated. These analytical designs, that is, nested case-control studies and case-cohort studies, are discussed in detail in chapter ▸Modern Epidemiological Study Designs of this handbook and will not be further considered here.

6.3.2 Description of Outcome Events in the Cohort

Cohort studies allow direct comparisons of exposed and unexposed subgroups and can provide measures of effects for various outcome events, such as different endpoints (morbidity, mortality, pre-morbidity) and/or different diseases. Nevertheless, analysis of cohort data requires reasonable care especially in the steps of data preprocessing for description and analysis. The often necessary change of perspective from persons at risk to person-time at risk needs special attention to ensure that unbiased results can be obtained. This subsection will refer mainly to disease incidence; however, other measures can, in principle, be treated in the same manner.

The results from a cohort study can be presented as shown in Table 6.1.

The easiest way to describe outcome events in a cohort is by counting the number of persons experiencing the event of interest and relating this number to the number of persons at risk in the cohort. Disease incidence, for example, can be described by the cumulative incidence or risk, which is calculated by dividing the number of incident cases by the number of persons at risk at baseline:

$$\widehat{\text{Risk}} = \text{number of incident cases}/\text{number of persons at risk}, \tag{6.1}$$

that can be calculated as

$$\widehat{\text{Risk}} = (a + c)/N \tag{6.2}$$

Table 6.1 2 × 2 table summarizing the results of a cohort study

		Second observe		
		Disease contracted	No disease	Total
First select	Exposed	a	b	$a + b = n_E$
	Non-exposed	c	d	$c + d = n_{\bar{E}}$
Total		$a + c$	$b + d$	$a + b + c + d = N$

and, accordingly for the exposed and unexposed study populations, as

$$\widehat{\text{Risk}}_E = a/(a+b) = a/n_E$$

$$\widehat{\text{Risk}}_{\bar{E}} = c/(c+d) = c/n_{\bar{E}}.$$

The cumulative incidence or risk is unit free and represents an individual risk of developing the disease. It is a proportion, not a rate, and it does not account for possible different periods of disease-free follow-up time of cohort members and assumes a fixed cohort. Thus, it is necessary always to specify the time period for incidence ascertainment for any cumulative incidence to avoid the fallacy of comparing cumulative incidence from different studies with different follow-up periods.

In cohort studies on acute diseases with short induction periods and a short time of follow-up, like outbreaks, the risk of disease can be estimated directly using the cumulative incidence, given a fixed cohort with fixed period of follow-up and a low fraction of dropouts. In cohort studies on chronic diseases with their long follow-up periods, however, the use of the cumulative incidence is not appropriate because usually disease-free follow-up periods differ among cohort members. In this case, occurrence of outcome events is preferably described by rates that represent the number of outcome events divided by the cumulated duration of event-free follow-up periods of all cohort members at risk. For further analysis, all rates presented in the following can be used to determine rate ratios and rate differences as described in chapter ▶Rates, Risks, Measures of Association and Impact of this handbook. Disease incidence can be expressed as incidence rate (I):

$$\hat{I} = \text{number of incident cases/cumulative person-time at risk,} \tag{6.3}$$

where each cohort member is contributing the time from entry into the study to either development of disease or end of follow-up to the denominator of the incidence rate, thus accounting for different times at risk of the cohort members to develop the disease. The incidence rate is sometimes called incidence density or hazard rate and should not be confused with cumulative incidence. Assuming equal person-time of follow-up of t, with t_E and $t_{\bar{E}}$ follow-up of exposed (E) and unexposed ($\bar{E}$) populations, (6.3) results in

$$\hat{I} = (a+c)/(N * t), \tag{6.4}$$

where $N * t$ denotes the total person-time at risk. Calculating the incidence rates separately for the exposed and unexposed study populations gives

$$\hat{I}_E = a/[(a+b) * t_E] = a/(n_E * t_E),$$

$$\hat{I}_{\bar{E}} = c/[(c+d) * t_{\bar{E}}] = c/(n_{\bar{E}} * t_{\bar{E}}).$$

Measures of risk and incidence of disease may provide important information regarding the public health burden of the outcome or disease of interest. Since incidence rates often vary considerably, for example, by age, sex, calendar year, and race, the calculation of specific incidence rates instead of crude incidence rates may be desirable. For this purpose, different strata (for one group variable) or cells (for two or more group variables) have to be defined over the group variables' range.

The individual contributions of the cohort members to numerator and denominator of the incidence rate have to be assigned to the respective stratum or cell. Usually, each cohort member will contribute to more than one stratum or cell as he/she moves through the cohort during follow-up. Age- and calendar-specific incidence rates can be approximated well enough on the base of calendar year data if more precise information on months and days is not available (see Breslow and Day 1987).

A simple example demonstrates the principle steps for the calculation of specific incidence rates for the age groups 30–39 years, 40–49 years, and 50–59 years. Table 6.2 shows the data of a fictitious cohort, for which we will calculate age-specific incidence rates. Since exact dates in terms of months and days are not available in our example, age and follow-up time will be approximated by full and half years. The contribution of the year at entry into the study and the year of diagnosis is approximated as half a year (see Fig. 6.2).

The cohort consists of 10 persons who were followed for 9–20 years resulting in a total of 155 person-years of follow-up, deaths, and dropouts accounted for the lacking 45 person-years. Three cases of the disease of interest occurred in the cohort during follow-up, resulting in a crude incidence rate of $3/135 = 0.022$ cases/person-year. The difference between the total of 155 observed person-years and the 135 person-years in the denominator of the incidence rate results from 20 years of cumulated follow-up time after diagnosis in the three cases. A useful general way in which to think of cohort data is to separate person-time at risk and person-time under observation.

Table 6.2 Data from a fictitious cohort

No.	Age at entry	Years of follow-up	Age at end of follow-up	Age at diagnosis	Person-years at risk
1	34	15	49		15
2	39	20	59	54	15
3	31	12	43		12
4	36	17	53	41	5
5	38	9	47		9
6	38	16	54	51	13
7	41	11	52		11
8	32	20	52		20
9	39	18	57		18
10	42	17	59		17
Total		155			135

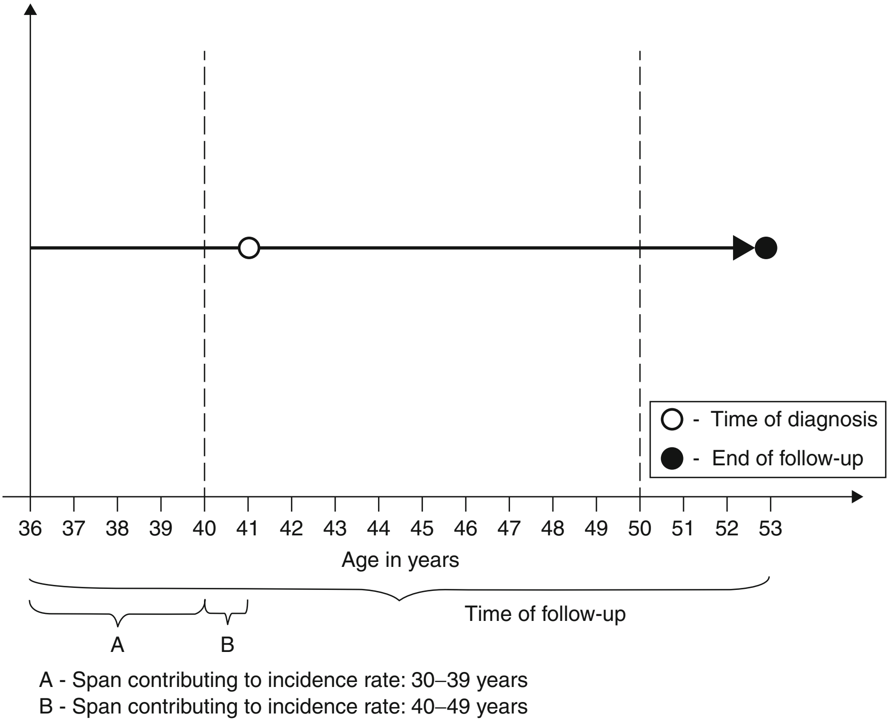


Fig. 6.2 Follow-up time of cohort member no. 4 of the fictitious cohort with respect to age

A subject is "at risk" at a given moment if the event of interest can happen. Thus, if a subject gets thyroid surgery, she/he is no longer at risk of getting a thyroid cancer. If on the other hand the event of interest were a pregnancy, a woman would not be "at risk" of becoming pregnant if she already is pregnant or during spells of abstinence. In this case, however, the woman is "at risk" again from the moment she desires another child. In the example above, a subject is no longer considered "at risk" after diagnosis of the disease. Of course, no subject is "at risk" from the moment of his/her death. Being at risk depends only on the endpoint studied.

On the contrary, being under observation (i.e., being followed up) depends on the precise definition of the cohort and the method of follow-up adopted in the study. A subject is under observation at a time t, if, were the event of interest to occur at this moment, it would be recorded. Thus, for example, if the cohort definition were "all subjects employed in a given factory with at least 1 year of employment," the follow-up would start only at the moment the subject satisfies this criterion. In this case, all the person-time in the first year must be ignored as if the event of interest occurred in this year; it would not satisfy the inclusion category. Similarly, a subject would be dropped from the follow-up at a time t if no information as to his/her disease status could be retrieved from time t on (e.g., the subject moves abroad), and the

Table 6.3 Age-specific incidence rates for fictitious cohort data

Age group	Incident cases	Disease-free follow-up time (years)	Age-specific incidence rate
30–39	0	$5.5 + 0.5 + 8.5 + 3.5 + 1.5 + 1.5 + 0 + 7.5 + 0.5 + 0 = 29$	$0/29 = 0$
40–49	1	$9.5 + 10 + 3.5 + \mathbf{1.5} + 7.5 + 10 + 8.5 + 10 + 10 + 7.5 = 78$	$1/78 = 0.013$
50–59	2	$0 + \mathbf{4.5} + 0 + \mathbf{0} + 0 + \mathbf{1.5} + 2.5 + 2.5 + 7.5 + 9.5 = 28$	$2/28 = 0.071$

subject is then considered “lost to follow-up” or “censored.” A subject contributes person-time to the study at any moment t if and only if at this moment he/she is “at risk” and “under observation.”

Coming back to the example, each incident case is assigned to the age group he/she belonged to until diagnosis. In the same manner, the disease-free time of follow-up of each cohort member is allocated to the three age groups yielding the age-specific incidence rates presented in Table 6.3.

Incidence rates are commonly rescaled, for example, to cases per 100,000 person-years underlining their reference to populations rather than to individuals. The crude incidence rate of 0.022 cases/person-year of the fictitious cohort, for example, would then be expressed as 2,222/100,000 person-years.

In Fig. 6.2, the follow-up time of cohort member no. 4 is depicted schematically with respect to age. The first three and a half years, denoted with A, of the 5 years of disease-free follow-up time (41 years at time of diagnosis – 36 years at entry into the study) are contributing to the denominator of the incidence rate of the first age group (30–39 years), and the next one and a half year, denoted with B, contribute to the numerator of the incidence rate of the second age group (40–49 years).

To quantify the frequency of exposure in the population under study, the prevalence of exposure may be considered:

$$\hat{P}_E \quad = (a + b)/N = n_E/N. \tag{6.5}$$

Note that this assumes that the exposures do not change in individuals over time. The various quantities presented here can be used to derive measures of association accordingly (see Sect. 6.3.4).

6.3.3 External Comparisons

One important form of validation of cohort studies is the ability to make comparisons with external data, preferably from the general population. Irrespective of the existence of internal comparison groups, external comparisons always give valuable insights by putting the cohort data in a broader context. For external comparisons, either age-, sex-, and calendar year-specific incidence or mortality rates or cumulative measures can be used. Standardized incidence rates can be calculated from specific incidence rates by weighting them with the age, sex, and calendar year distribution of the external comparison data (direct standardization). However, cumulative measures have to be interpreted cautiously since they can mask underlying differences in specific disease patterns, such as an unusually high incidence rate among younger persons in the cohort. With d_i denoting the number of cases in the age group i, n_i denoting the disease-free person-years accumulated in the age group i, and w_i denoting the proportion of persons in the age group i in the standard population, the directly age-standardized incidence rate $\hat{I}_W$ is calculated as

$$\hat{I}_W = \sum_{i=1}^{I} w_i d_i / n_i . \tag{6.6}$$

Indirectly standardized measures requiring morbidity or mortality rates of the standard population are the standardized morbidity or incidence ratio (*SIR*) and the standardized mortality ratio (*SMR*). Since morbidity data are not routinely available in most countries, the standardized mortality ratio is used much more frequently. The *SMR* compares the observed numbers of deaths in the cohort with the expected numbers, given the age structure of the cohort and the age-specific mortality rates λ_i of a reference population. With d_i denoting the number of deaths in the age group and n_i denoting the person-years accumulated in the age group, the *SMR* is estimated as

$$\widehat{SMR} = \sum_{i=1}^{I} d_i \Big/ \sum_{i=1}^{I} n_i \lambda_i , \tag{6.7}$$

where $\sum\limits_{i=1}^{I} d_i$ represents the total number of observed deaths in the cohort under investigation and $\sum\limits_{i=1}^{I} n_i \lambda_i$ the expected number of deaths that are obtained by applying age-specific incidence rates of the reference population to the cohort under investigation. An *SMR* above 1 indicates a larger mortality in the cohort, and an *SMR* below 1 a smaller mortality in the cohort compared to that of the reference population. Statistical testing of a single *SMR* can be done with a simple χ^2–test (observed vs. expected) with one degree of freedom. Assuming that the number of observed cases $D = \sum\limits_{i=1}^{I} d_i$ follows a Poisson distribution with expectation $\gamma = E(D)$, confidence limits for the *SMR* ($\widehat{SMR}_L$, $\widehat{SMR}_U$) can be obtained by finding confidence limits $\hat{\gamma}_L$, $\hat{\gamma}_U$ for the number of observed cases:

$$\widehat{SMR}_L = \hat{\gamma}_L / \sum_{i=1}^{I} n_i \lambda_i \text{ and } \widehat{SMR}_U = \hat{\gamma}_U / \sum_{i=1}^{I} n_i \lambda_i . \tag{6.8}$$

The confidence limits for γ can be determined as

$$\hat{\gamma}_L = (1/2)\chi^2_{2D,\alpha/2} \text{ and } \hat{\gamma}_U = (1/2)\chi^2_{2(D+1),1-\alpha/2}, \tag{6.9}$$

where $\chi^2_{2D,\alpha/2}$ denotes the $100(\alpha/2)$th percentile of the χ^2-distribution with $2D$ degrees of freedom and $\chi^2_{2(D+1),1-\alpha/2}$ denotes the $100(1-\alpha/2)$th percentile of the χ^2-distribution with $2(D+1)$ degrees of freedom (Sahai and Khurshid 1996).

If the age-specific rates of the reference population are just estimations of the true rates, as is often the case with morbidity data, calculation of confidence intervals for *SMR* can be performed by the method described in Silcocks (1994). A method for estimating *SMR* where information on vital status is complete but information on cause of death is partly missing, as may be the case in historical cohort studies, can be found in Rittgen and Becker (2000).

Comparison of rates by direct standardization has poor statistical properties, especially due to large variances of age-specific rates in small cohorts. Therefore, indirect standardization is usually preferred (see also chapter ►Rates, Risks, Measures of Association and Impact of this handbook).

6.3.4 Internal Versus External Comparisons

In the previous section, the event rate (morbidity or mortality) of a cohort is compared to the rates of an external population. This is done by comparing the observed number of deaths in the cohort with the expected numbers, given the age structure of the cohort and the age-specific mortality rates λ_i of a reference population. The ratio of observed to expected (the *SMR*) is then interpreted as a rate ratio between the cohort and the general population taken as the reference.

If the cohort is set up for investigating a specific risk factor, as would be the case in an occupational cohort, one can be tempted to interpret the *SMR* as a risk ratio due to the risk factor under investigation. However, this interpretation would only be valid if the cohort were comparable to the general population for all factors except for the risk factor under investigation. This is obviously only rarely the case. The general population consists of all subjects including the very ill and very poor, which would rarely be included in the same proportion in a cohort. Thus, the mortality in the general population is usually higher than in any (unexposed) cohort. In occupational cohorts, this phenomenon has been termed the "*Healthy Worker Effect*" (see, e.g., Li and Sung 1999; Goldberg and Luce 2001). Other factors, like regional differences, owing to social, behavioral, nutritional, and environmental factors, might cause the mortality of a regionally based cohort to be different from a nationwide general population.

In summary, the *SMR* is a potentially biased estimate of the effect of any risk factor. This bias can be reduced by choosing a reference population which is as similar as possible (except for the risk factor of interest) to the cohort under investigation or by comparing the cohort to another reference cohort. In the latter case, however, the computation of the confidence interval of the *SMR* is no longer valid as it assumes that, because of the large number of subjects in the reference population, the disease rates and hence the expected numbers are observed without any sampling error. Then the only statistically valid methods are those presented in the following section, although the confidence intervals of the risk ratio become wider. The choice between an external comparison and an internal comparison is thus the choice between accepting a (often small) bias and accepting a larger variance, which implies a lower power. Such a choice can only be made in the context of each study, and, if possible, both approaches should be tried. Finally, methods have been proposed including external reference rates to stabilize internal comparisons (e.g., Breslow and Day 1987, p. 151) that might be used as a reasonable compromise.

6.3.5 Summary Effects of Exposure

The main goal of cohort studies is to compare morbidity and/or mortality in exposed and non-exposed subjects or between different exposure groups of the cohort and to investigate dose-effect relationships between exposure and disease. If the exposure is constant and can be determined at entry into the cohort, internal comparisons can be performed by calculating specific incidence rates for each exposure category separately as if each group were a separate cohort. Cumulative rates can be used, again provided the subgroups do not differ in important determinants of disease, for example, age.

In the simple case of a single dichotomous exposure, several measures of association of exposure with disease can be estimated from results provided by a cohort study (see Table 6.1). In the following, the most important ones will be briefly introduced. A detailed discussion of their properties and examples for their calculation can be found in chapter ▸Rates, Risks, Measures of Association and Impact.

Perhaps the most popular measure of association is the risk ratio (*RR*), also known as relative risk that compares the experience of exposed and unexposed populations. With the notation given in Table 6.1 and the risks for the exposed and unexposed subjects calculated according to (6.2), it can be estimated as

$$\widehat{RR} = \widehat{\text{Risk}}_E / \widehat{\text{Risk}}_{\bar{E}} = [a/(a+b)]/[c/(c+d)] = (a/n_E)/(c/n_{\bar{E}}). \qquad (6.10)$$

The incidence ratio (*IR*) compares the incidence rates in the exposed and unexposed study populations. According to (6.4), its estimator is given as

$$\widehat{IR} = \hat{I}_E/\hat{I}_{\bar{E}} = \{a/[(a+b)*t_E]\}/\{c/[(c+d)*t_{\bar{E}}]\} = [a/(n_E * t_E)]/[c/n_{\bar{E}} * t_{\bar{E}})]. \qquad (6.11)$$

The *RR* and *IR* provide estimates of the relative strength of the association between the exposure of interest and the outcome or disease of interest.

The absolute difference in risk (*AR*) between the exposed and unexposed groups provides an estimate of the impact of the exposure on the risk of disease in absolute terms. This measure is not to be confused with the absolute risk, which is the absolute probability that a disease-free individual will develop a given disease over a specific time interval (Benichou 1998). Using the above formulas for the risks among exposed and unexposed, it can be obtained from a cohort study as

$$\widehat{AR} = \widehat{\text{Risk}}_E - \widehat{\text{Risk}}_{\bar{E}} = [a/(a+b)] - [c/(c+d)] = a/n_E - c/n_{\bar{E}}. \quad (6.12)$$

Based on the concept of attributable risk, several other measures can be derived, though care must be used in their computation and interpretation (Greenland and Robins 1988, and see chapter ▸Rates, Risks, Measures of Association and Impact of this handbook). The attributable fraction (AF) can be interpreted as the proportion of risk due to exposure in exposed individuals. It may be useful for quantifying the degree to which risk can be reduced at the individual level if the exposure (and its effects) can be eliminated. It may, therefore, be a sensible measure for counseling individuals:

$$\begin{aligned}\widehat{AF} &= \widehat{AR}/\widehat{\text{Risk}}_E = \{[a/(a+b)] - [c/(c+d)]\}/[a/(a+b)]\\ &= (a/n_E - c/n_{\bar{E}})/(a/n_E).\end{aligned} \quad (6.13)$$

The population attributable risk (PAR) reflects the absolute level of risk of the outcome in the population due to the exposure. It can be used to estimate the public health impact, in absolute terms, of elimination of the exposure, at least with respect to the outcome of interest. Based on the attributable risk and the prevalence of exposure (see Eq. 6.5), it is given as

$$\begin{aligned}\widehat{PAR} &= \widehat{AR}/\hat{P}_E = \{[a/(a+b)] - [c/(c+d)]\}/[(a+b)/N]\\ &= (a/n_E - c/n_{\bar{E}})/(n_E/N).\end{aligned} \quad (6.14)$$

The last measure to be mentioned here may be used to estimate the proportion of all events of interest that could be prevented in the overall population if the exposure (and its effects) can be eliminated. The population attributable fraction (PAF) is defined as the proportion of all events of interest that occur in the population due to the exposure:

$$\widehat{PAF} = \widehat{PAR}/\widehat{\text{Risk}} = (a/n_E - c/n_{\bar{E}})/\{(n_E/N)[(a+c)/N]\}. \quad (6.15)$$

6.3.6 Internal Modeling of the Effects of Exposure

The situation is more complicated if cohort members continuously add exposure over follow-up time. Simple categorization on the basis of cumulative exposure would lead to biased results. Person-years accumulated shortly after entry into the study of cohort members with high cumulative exposure would wrongly be allocated to a high exposure category, although the cumulative exposure at that time point was still low for these cohort members, resulting in underestimation of high exposures and overestimation of low exposures. Therefore, the disease-free

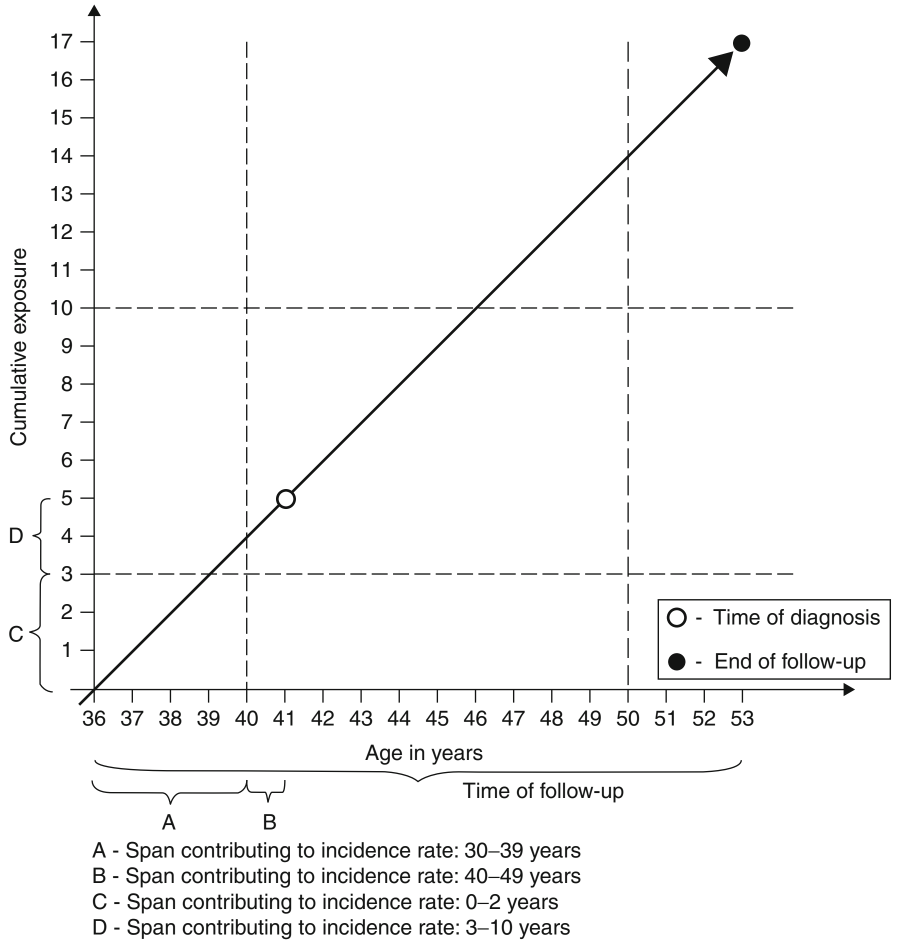


Fig. 6.3 Follow-up time of cohort member no. 4 of the fictitious cohort with respect to age and cumulative exposure

person-time of each subject has to be subdivided and assigned to the respective age- and sex-specific exposure category the cohort member belongs to as he or she moves through the cohort, meaning that most cohort members contribute to different age-exposure categories. In the same manner, the incident cases have to be assigned to the categories where they occurred.

In Fig. 6.3, the follow-up time of cohort member no. 4 is again depicted schematically, this time with respect to age and cumulative exposure assuming that the exposure starts at the beginning of the follow-up and that it is constant over time. For age- and exposure-specific incidence rates, the disease-free-follow up time is assigned to the groups according to the squares in the figure that are defined by the categorization of the group variables, resulting in a contribution of cohort member no. 4 of two and a half years to the denominator of the incidence

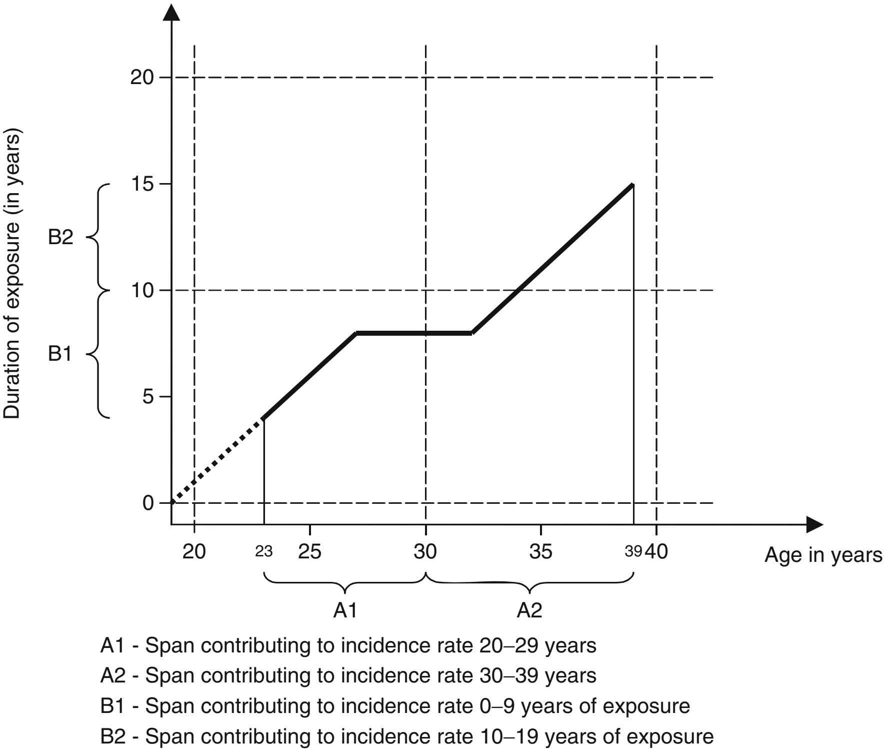


Fig. 6.4 Person-time classification with varying duration of exposure

rate of category A × C (30–39 years of age and <3 units of cumulative exposure), 1 year to the denominator of the incidence rate of category A × D (30–39 years of age and 3 – <10 units of cumulative exposure), and one and a half year to the denominator of the incidence rate of category B × D (40–49 years of age and 3 – <10 units of cumulative exposure). The case itself contributes to the numerator of the incidence rate of category B × D, since this is the category in which he/she was diagnosed.

This procedure can be extended in several ways. The exposure may have started before beginning of follow-up or may start later. It can vary over time, and it can even vary from individual to individual or can be lagged to account for induction time. Several measures of exposure (e.g., time since first exposure and/or confounders) can be considered simultaneously, and possible confounders can be included in the analyses as additional variables. Figures 6.4–6.6 illustrate some of these features. For simplicity, no half years are considered in these examples.

In Fig. 6.4, a subject is followed up from age 23 but has been exposed from age 19 on; he/she is exposed until age 27 followed by an unexposed 5-year period. He/she is again exposed until age 39 at which time his/her person-time

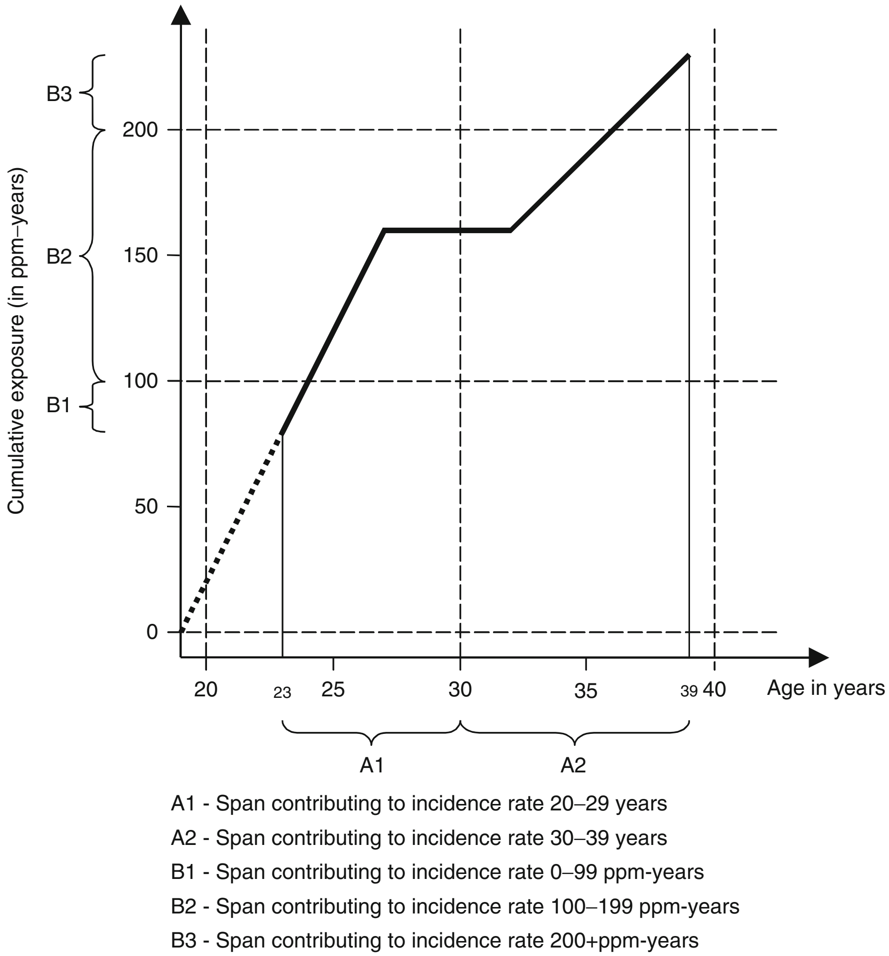


Fig. 6.5 Person-time classification with varying cumulative exposure

at risk ceases either because of disease diagnosis or because of end of follow-up. This subject would contribute 7 years (from age 23 to age 30) to the A1 × B1 group (20–29 years of age, 0–10 years exposure), 4 years (from age 30 to 34) to the A2 × B1 (30–39 years of age, 0–10 years of exposure), and 5 years (from age 34 to 39) to A2 × B2 (30–39 years of age, 10–19 years of exposure).

Figure 6.5 presents the same subject assuming that the first exposure spell was twice as intensive (e.g., 20 ppm of a given chemical) than the second exposure (10 ppm). The units of cumulative exposure on the y-axis are now ppm-years. The subject would contribute 1 year to group A1 × B1 (his cumulative exposure is then 100 ppm-years), then 6 years to group A1 × B2 (at age 30, his cumulative exposure

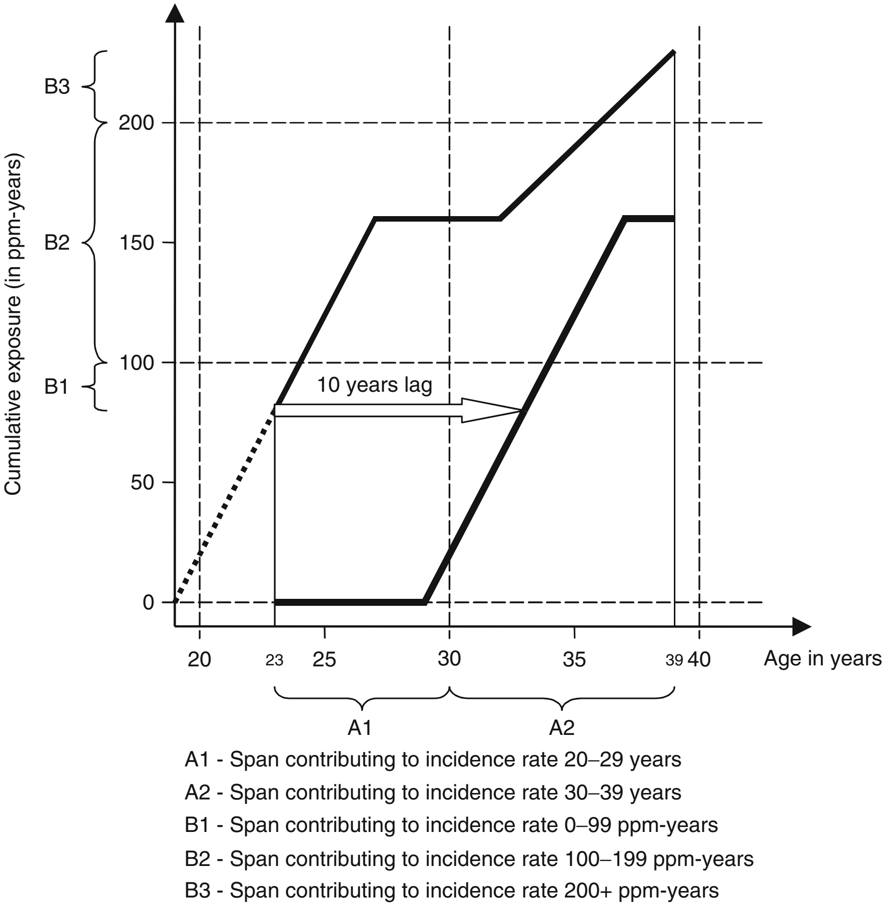


Fig. 6.6 Person-time classification with varying lagged cumulative exposure

is 160 ppm-years), then 6 years (from age 30 to age 36 at which he reaches 200 ppm-years) in group A2 × B2 and finally 3 years in group A2 × B3.

Figure 6.6 considers the same subject again, but this time, the exposure is lagged by 10 years, say, to account for disease induction time. The first period would then be a non-exposed period. The rationale is that, were the disease to occur in these first 10 years, it would not be attributable to exposure. Applying the same rationale as before, the subject would stay 6 years in 0 × A1, then the subject would contribute 1 year in group B1 × A1, 4 years in group B1 × A2, and finally 5 years in group B2 × A2; the lagged cumulative exposure at end of follow up (i.e., at age 39) is 160 ppm-years.

Another exposure can occur during the follow-up, for example, the preceding subject starts smoking at age 25. In this case, a further splitting of the time periods would be done separating periods in which the subject was a non-smoker and periods in which he/she smoked.

This splitting of person-time into age and exposure groups must be done for each subject of the cohort and gets more complex with a growing number of group variables. Specialized software packages exist (e.g., Coleman et al. 1986) to perform these computations, but they are usually limited in the complexity they can handle. Interestingly, these restrictions do not apply to some more general packages as Stata (version 7 or later – StataCorp 2001) or Epicure (Preston et al. 1993) in which the statistical modeling procedures of such data are included. The end result of the calculations carried out in these packages can then be presented as a data table with each line corresponding to a separate combination of age and exposure classes (other classifications like calendar periods might also be included) and containing the following variables: the value of each age and exposure group, the number of person-years n_i accumulated in this category over the entire cohort, and the number d_i of events of interest falling in this category.

In cohort studies, the standard model for analyzing such data is the Poisson model which is a statistical model of the disease rates. Basically, the Poisson model assumes that the number of events d_i in each category i (combination of age category j and the kth combination of exposure variables) follows a Poisson distribution with parameter $n_i \lambda_i$. The standard (multiplicative) model would then assume that

$$\ln(\lambda_i) = \alpha_j + \beta_k, \tag{6.16}$$

where λ_i are the unknown true disease rates, the α_j are nuisance parameters specifying the effects of age and (possibly) other stratification variables such as calendar periods, and β_k the parameters that describe the effects of primary interest. As usual, in regression models, $\beta_0 = 0$ would be a baseline category. $\exp(\beta_k)$ is then an estimate, adjusted on the nuisance parameters, of the relative risk of the kth exposure category versus the baseline category assuming absence of interaction between exposure. The full modeling strategy of the Poisson regression is beyond the scope of this chapter but is not different from any regression modeling (see chapters ▶Regression Methods for Epidemiological Analysis and ▶Survival Analysis of this handbook). A comprehensive account of Poisson modeling is given by Breslow and Day (1987, Chap. 4).

An alternative way of analyzing event history data (another denomination of cohort data focused on events) is by using Cox's proportional hazards regression model. This model acknowledges that the categorization of continuous data always implies a loss of information and therefore generally a loss in statistical power. Moreover, there is no need to explicitly estimate the effects of nuisance parameters if it can be avoided.

The first step in proportional hazard modeling is the choice of one of the time variables considered. This basic time variable can either be age, as was implicit at the beginning of this chapter, or in some settings, the calendar time or even the time since the beginning of follow-up. Once this time variable has been fixed, its effects can be estimated non-parametrically.

The key idea of Cox's regression is that no information is lost when considering only the time points t_i at which an event of interest occurs. At each such time point, a "risk set" is established including all members of the cohort contributing person-time (at risk and under observation) at this time point. If one wants to use a Cox model, the first step is thus to identify all risk sets. Then, one must obtain the value at each time t_i of all variables to be included in the model for all members of the corresponding risk set. The statistical analysis is then similar (in fact, the same software can be used) to a conditional logistic regression analysis, in which the matching variable is the indicator of the risk set. As in the logistic regression, the exposure at time t_i of the case, that is, the subject experiencing the event at time t_i, and the exposure at time t_i of the other members of the risk set are compared. Again, the full modeling strategy of the Cox's proportional hazards regression model and its various extensions are beyond the scope of this chapter (see chapters ▸Regression Methods for Epidemiological Analysis and ▸Survival Analysis of this handbook). A comprehensive though somewhat dated account of this model is given by Breslow and Day (1987, Chap. 5). As for Poisson models, both Stata and Epicure provide easy to use software, but once the risk sets and the corresponding exposure variables have been computed for each risk set, any logistic regression package (e.g., Proc PHREG in SAS) can be used.

6.4 Key Concerns in Cohort Studies

6.4.1 Selection of the Study Population

If external comparisons are to be made, usually, vital statistics data of the general population or data derived from national disease registries are used as a reference for the calculation of expected cases. However, they can only be regarded as valid for deriving an expectation of mortality and disease rates if the cohort under investigation is a representative sample of the general population. Indeed, many cohorts are convenience samples, derived from a group that happens to be accessible. Representative cohorts can, for example, be derived from national censuses, utilizing the data collected for the specific census. Obtaining access to census data is generally not easy, however, since most censuses guarantee confidentiality to participants. Exceptions to that rule are, for example, a Swedish occupational census-based sample and a 10% sample of the Canadian labor force, derived from data collected on Canadians having a social insurance number required for all who are employed in an active occupation (Howe and Lindsay 1983). These types of population samples are very valuable because subsets among them chosen for specific analysis can be regarded as comparable to the general population apart from the characteristics that caused them to enter, or be selected for, that subset.

Occupational cohorts (cf. chapter ▸Environmental Epidemiology of this handbook) are usually identified by company files or sometimes by workers' union files. Access to these cohorts is usually granted if the company or union is interested in determining whether a suspected increase in disease rates has occurred or there is concern that exposure to a potential hazard bears an increased risk of disease. For example, many carcinogens have been confirmed to be hazardous to humans, often after evidence first derived from animal studies, by investigations of specific cohorts (Tomatis et al. 1990). This mechanism is still being used, as exhibited by a tri-utility study of electrical and magnetic field exposures (Theriault et al. 1994) and a study of Motorola employees on the potential risks of exposure to radiofrequency fields (Morgan et al. 2000). It is very helpful if employment records indicate exposure to specific agents. This is the case when routine measurements are taken for safety reasons, as for most workers exposed to radiation. In their absence, estimation of exposures may be required, as discussed further below.

So-called multipurpose cohorts identified for study, however, have to be recruited by some mechanism that provides the opportunity for potential subjects to volunteer. For example, much has been learnt from an ongoing study of American nurses, who were given the opportunity to volunteer for the study by completing a questionnaire of dietary and other lifestyle factors (Willett et al. 1992). Similar studies were initiated in Canada by providing self-administered questionnaires to women already participating in a mammography screening trial (Howe et al. 1991) and in Sweden by approaching women who participated in a routine mammography screening program (Wolk et al. 1998). In Europe, a large multicenter cohort study was initiated in 10 countries using different approaches (Riboli and Kaaks 1997). Some used population registers as the basis for mailing invitations to participate. The response proportions were good in most countries, but still tended to include more health conscious and more highly educated people than the general population as is often the case in volunteer studies (cf. chapter ▸Epidemiological Field Work in Population-Based Studies of this handbook).

Another recent feature of cohort studies has been the attempt to bring many together and analyze them almost as a multicenter study to enable the investigators to identify risks which none of them individually were capable of demonstrating. The Pooling Project is a case in point, originally funded to evaluate further uncertain associations between diet and breast cancer; it has proven a very useful source of additional knowledge because of the ability of cohort studies to identify multiple endpoints (Smith-Warner et al. 2006). Thus, it has already been extended to lung cancer (Smith-Warner et al. 2003), with findings similar to the EPIC study (Miller et al. 2004), and other diseases have followed (e.g., Cho et al. 2004; Koushik et al. 2006).

When a truly representative cohort cannot be obtained, because the mechanism used involves the opportunity to volunteer and to refuse to participate, comparisons with the general population in terms of mortality and disease rates may not be valid. Thus, the cohort study may lack external validity. However, provided that the recruitment mechanism is unbiased with regard to the exposure of interest, and the data obtained on exposure enables the investigators to stratify their population into exposed and unexposed subgroups, the estimation of the association between the exposure and the outcome will be valid (internal validity).

Tables 6.4 and 6.5 demonstrate the effects of different participation patterns (selection) on estimates that can be obtained from cohort studies. In the presence

Table 6.4 Effects of a fair sampling process on the measures of disease occurrence and association

Target Population

Exposed	Disease: Yes	No	Total
Yes	40	160	200
No	60	740	800
Total	100	900	1,000

→ Selection Weights

50	50
50	50

→ Study Population

Exposed	Disease: Yes	No	Total
Yes	20	80	100
No	30	370	400
Total	50	450	500

Target Population		Study Population
10%	Prevalence of Disease	10%
20%	Prevalence of Risk Factor	20%
2.67	Relative Risk	2.67
3.08	Odds Ratio	3.08
125/1,000	Attributable Risk	125/1,000

Table 6.5 Effects of oversampling of exposed individuals on the measures of disease occurrence and association (positive association between exposure and outcome)

Target Population

Exposed	Disease: Yes	No	Total
Yes	40	160	200
No	60	740	800
Total	100	900	1,000

→ Selection Weights

100	100
10	10

→ Study Population

Exposed	Disease: Yes	No	Total
Yes	40	160	200
No	6	74	80
Total	46	234	280

Target Population		Study Population
10%	Prevalence of Disease	16%
20%	Prevalence of Risk Factor	71%
2.67	Relative Risk	2.67
3.08	Odds Ratio	3.08
125/1,000	Attributable Risk	125/1,000

of a fair sample, all of the measures of disease occurrence and association will be unbiased (Table 6.4). In the presence of overrepresentation of exposed persons (Table 6.5), the prevalence of the exposure will be overestimated and the risk of the outcome will be over- or underestimated depending on whether the exposure is positively or negatively associated with disease. Nevertheless, the estimates of the relative risk and the attributable risk will be unbiased. Since the estimate of the prevalence of exposure is biased, estimates of the public health impact will be biased. Other participation patterns that can theoretically introduce selection bias including overrepresentation of diseased individuals and participation rates that differ by both exposure and disease status are unlikely to affect cohort studies due to the customary exclusion of persons with the outcome of interest at baseline. This assurance is only relative, relying on the degree to which persons with prevalent disease can be excluded from the cohort. In general, selection bias can be minimized by avoiding the use of volunteers (or using volunteers exclusively) and by minimizing non-participation. The potential for selection bias can be assessed by evaluating non-participants for study characteristics, if possible.

6.4.2 Exposure and Confounders in Cohort Studies

As already indicated, some cohorts will have exposure data readily available, especially those derived from occupational groups where exposure was routinely collected for safety monitoring purposes. It is the strength of such cohorts that they offer the possibility to report the exposure before the disease occurs. However, for population-based cohorts, the investigators will have to collect data specifically for the study or refine existing data.

Because most cohorts will be very large, the collection of exposure data is not a simple task. If exposure data is to be collected by questionnaires, the scale of the effort required will generally mean that neither personal nor telephone interviews are feasible, as would normally be planned for case-control studies. This means that the exposure data will generally be collected by mailed self-administered questionnaires, often linked to the recruitment mechanism of the cohort, with response to the questionnaire qualifying the individual for inclusion in the study. Inevitably, the amount of data that can be collected by self-administered questionnaire is limited. The degree of detail for a given variable that can be obtained by such instruments is also restricted (cf. chapter ▸Exposure Assessment of this handbook), so that in addition to the problems of the ability of the respondent to recall accurately the exposure he/she has experienced, the data will be potentially subjected to major misclassification.

The extent of misclassification in cohort studies has only recently been appreciated, probably explaining the fact that the results of many cohort studies, especially when diet was the exposure of interest, have been negative (Day and Ferrari 2002). Thus, although many of the questionnaires used in cohort studies have been subject of validation studies, and correlation with other assessment methods seemed reasonable, these validation studies have served to reassure the investigators but probably have not protected them from reporting negative or very weak results. Even for smoking, the information obtained in cohort studies cannot be regarded as precise as investigators would have wished.

Misclassification of exposure can be differential or non-differential with respect to the outcome of interest, that is, the degree of misclassification of the exposure can

Table 6.6 Non-differential misclassification of exposure

	True Cohort (no error)			Observed Cohort (error)		
Cigarettes	MI	No MI	Total	MI	No MI	Total
Yes	60	300	360	54	270	324
No	30	330	360	36	360	396
Total	90	630	720	90	630	720
Rate Ratio	2.00			1.83		

differ, or not, by outcome status. In cohort studies, non-differential misclassification is the more typical form of misclassification due to the customary exclusion of persons with prevalent disease at baseline. It is unlikely that the measurement of exposure at baseline will be influenced by the development of an outcome sometime in the future. Differential misclassification is potentially a much greater problem in case-control and cross-sectional studies. Non-differential misclassification always introduces a bias toward a null finding (a finding of no association) if the exposure status is dichotomized, whereas differential misclassification can introduce a less predictable bias. Table 6.6 shows the impact of a 10% non-differential error rate in classifying smokers, in a cohort study in which occurrence of myocardial infarction (MI) was the observed endpoint. In this example, 90% of exposed individuals were correctly classified regarding exposure and 100% of unexposed individuals were correctly classified. Assuming a true relative risk of 2.0, the observed relative risk would be 1.8. With greater degrees of misclassification, the bias toward the null would increase. This bias can be minimized through the use of standardized and validated procedures for exposure assessment. It should be noted, however, that the bias due to non-differential misclassification can be toward or away from the null if the exposure has more than two categories.

Another issue that affects cohort studies differently than case-control studies is the effect of change in exposure with time. In case-control studies, detailed exposure biographies that include changes in exposure patterns, for example, change in intensity of smoking, or cessation of smoking, or even measures taken to affect dietary change, can be retrieved using just one survey, with the problem of uncertainty, and possibly differential error, in recall. The concurrent cohort study with its prospective data collecting does offer the possibility of assessing changes in exposure while they happen. To assess changes in exposure patterns, a mechanism has, however, to be set up specifically, for example, by readministering the questionnaire on a regular basis. This could be done as part of the follow-up mechanism adopted, though some loss to follow-up will be inevitable. An alternative to incorporating this new information into the analysis is shown in the Nurses Health Study (Willett et al. 1992). The follow-up period with regard to the time from the first exposure information to the second was used as a separate cohort from the follow-up period subsequent to the second exposure information. This is justifiable as blocks of person-time in different periods are statistically independent, regardless of the extent they are derived from the same people (Rothman and Greenland 1998). However, usually cohorts are analyzed with regard to the exposure determined at baseline, and although that may seem distant from the period when many endpoints are determined, for those with a long induction period from exposure to outcome, as for many cancers, this has not always been regarded as a major disadvantage.

Exposure assessment by questionnaires always depends on subjects' accuracy of recall and their willingness to participate, and many efforts have been made to introduce more objective measures of exposure determination. For radiation exposure, cohorts with occupations that require wearing film badges provide cumulative and, in some instances, peak measurements of exposure. For uranium and other hard rock miners, measures of the radiation exposure in mines were often made for safety reasons to limit the length of exposure of those at risk, and these measurements can be assigned to the job history of the individual.

However, in many instances, exposure has to be estimated simply from the type of occupation at a certain time since no further information is available and misclassification of exposure assessment cannot be avoided. In occupational studies, attempts have been made to refine exposure assessment by developing a job exposure matrix (cf. chapters ▶Exposure Assessment and ▶Occupational Epidemiology of this handbook). Often, using data from hygiene assessments performed in the past, a matrix can be constructed with the different job tasks in the rows and columns indicating the probability and/or intensity of exposure within that job to the agents (chemical or physical) of interest. The approach was, for example, used in a study of electrical and magnetic field exposures in electric utility workers in Canada and France (Theriault et al. 1994). Extension of the work upon a sample of workers wearing portable electric and magnetic field exposure meters and using historical data of electrical usage in the province enabled the investigators to identify strong associations of leukemia and non-Hodgkin's lymphoma risk with high electric field exposure (Miller et al. 1996; Villeneuve et al. 1998).

Another source of exposure data collected in cohort studies is gained from biological material of the cohort members. Historically, rather simple parameters were under study, like blood pressure or cholesterol levels, derived from blood samples that were collected in the framework of large cohort and intervention studies on cardiovascular disease. Now, there is increasing interest in the study of disease etiology by biomarkers of exposure and/or of genetic factors, as, for example, in the European Prospective Investigation of Diet and Cancer (EPIC) (Riboli and Kaaks 1997).

The findings of cohort studies regarding the effects of exposure can be strengthened if it is possible to evaluate a dose-response relationship. This requires the assessment of intensity of exposure that can be quantified as peak, average, or cumulative exposure. Sometimes duration of exposure is used as a surrogate for cumulative exposure. However, using duration in this way is problematic if the exposure is associated with an early, perhaps toxic, effect. Then it could be anticipated that these workers would tend to change their employment and could not cumulate long durations of exposure. If such workers represented a particularly susceptible subgroup, perhaps for genetic reasons, it is possible that in this subgroup, a relatively brief exposure results in the same incidence of disease than in subgroups with a longer duration of exposure that are less susceptible. The absence of a dose-response relationship without appropriate statistical control for the genetic background might then be incorrectly interpreted as indicator that the exposure is not causal for the disease (Blair and Stewart 1992).

The treatment of potential confounding factors is the major challenge of the analysis of cohort studies. This is in part because the basic data set may not contain information on all relevant confounders, particularly not in historical cohort studies but also because the data available on confounders may not be assessed with sufficient precision to take account of their effect. An example is the possible confounding effect of cigarette smoking with fruit and vegetable consumption and lung cancer. Although two large cohort studies (one multicenter and one the result of a pooled analysis) which fully adjusted for the effects of cigarette smoking in the opinion of the investigators were available (Miller et al. 2004; Smith-Warner et al. 2003), a working group of the International Agency for Research on Cancer (IARC) was not convinced that there was not residual confounding of fruit consumption by smoking with lung cancer and therefore judged the evidence to be limited rather than sufficient (IARC 2003).

6.4.3 Determining Outcome Events

A limiting factor for cohort studies is that most diseases are relatively rare, with rates determined in the population per 100,000 persons. Therefore, to accrue sufficient cases of the disease, the size of the cohort has to be large and/or the follow-up time has to be long. Another factor affecting the length of follow-up relates to the long induction period from the beginning of many exposures to the occurrence of disease. For many cancers, for example, the induction period exceeds 10, often 20, years. One example for the importance of a long enough follow-up period is the British Doctors' Study that showed much higher lung cancer risks of cigarette smoking after 40 years of follow-up than in the 10- and 20-year reports of this study (Doll et al. 1994b). The reason for this was a dominant effect of duration of smoking compared to intensity of exposure on the risk of lung cancer (see also Flanders et al. 2003). It seems probable that this is not the only example of this phenomenon – it may particularly affect exposures with a long induction period from initiation of exposure to effect. The possibility of such an effect should encourage investigators to maintain the follow-up of well-documented cohorts for as long as proves feasible, and granting agencies will agree to provide the necessary funds. If grants are limited, it may be useful to store the necessary data and extend the follow-up after a certain time lapse. It is unusual for cohort studies to start from the first exposure and the possible initiation of disease, covering the whole spectrum of exposure in a subject's lifetime. Attempts have to be made to determine or to estimate past exposure, with all the error and potential misclassification of such inquiries. Nevertheless, a major advantage of cohort studies over case-control studies is that exposure is determined prior to the diagnosis of disease, thus avoiding a major bias of concern in case-control studies, that is, the information (recall) bias.

As already indicated, the follow-up of cohorts enables multiple endpoints to be determined, for example, different types of cardiovascular disease and/or different cancer sites. In determining endpoints in cohort studies, it is essential that ascertainment bias is avoided. Ascertainment bias relates to the possibility that the surveillance of cohort members, by virtue of the fact that they are in a study, may result in greater efforts to make a diagnosis than would occur in the general population. Special surveillance mechanisms in a cohort study are valid if internal comparisons of exposed versus unexposed within the cohort are planned but would invalidate external comparisons with general population data. Orencia et al. (1995) provided an example of this bias in a non-concurrent cohort study examining the association of mitral valve prolapse (MVP) with stroke. Using the database of the Mayo Clinic, they assembled a cohort of persons with MVP, followed them for the occurrence of stroke, and compared the rate of stroke with the rate in the general population of Olmsted County, Minnesota. The overall standardized mortality ratio was 2.1, indicating a risk of stroke twice of that of the general population. However, Orencia et al. noted that MVP can be diagnosed by auscultation or as a serendipitous finding during an echocardiogram conducted for other medical reasons (e.g., following myocardial infarction, chronic heart failure, atrial fibrillation) often associated with risk of stroke. When the cohort was further subdivided according to method of diagnosis, the auscultatory group demonstrated no increase in risk. The increased risk was confined to the group identified serendipitously during a cardiac evaluation motivated by other medical concerns associated with risk of stroke.

In some cohort studies, annual or less frequent contact by mail, generally with the cohort member directly or sometimes with his or her designated physician, will identify the probable occurrence of a study endpoint or death from a cause unrelated to the disease of interest. However, these processes are costly and also pose the risk of losing an increasing proportion of cohort members with time. Further, if the participant has died, family members may not always be willing to collaborate in providing the required information. Hence, in many studies, other mechanisms are used for follow-up and indeed may have to be used also for subjects lost if the basic mechanism of follow-up is by mail. Losses to follow-up lead to a loss of power due to the resultant loss of sample size and can introduce bias in a manner similar to the selection processes described previously. Losses that do not differ by either exposure or disease status result in a picture similar to that shown in Table 6.4, that is, no bias but a loss of power. Losses that differ by exposure (but not outcome) status introduce the same bias as that described in Table 6.5. More problematic are losses that differ by outcome status (Table 6.7) and those that differ by both exposure and outcome status (Table 6.8). In these situations, estimates of the relative risk may be biased in unpredictable directions.

Apart from special surveillance mechanisms, including screening for the disease of interest, there are many sources of routinely collected data for endpoints in cohort studies. These include medical records of physicians, health maintenance organizations and hospitals, vital statistics systems, and disease registries. The process to

Table 6.7 Effects of losses to follow-up that differ by outcome status on estimates of disease occurrence and association

Target Population

Exposed	Disease Yes	No	Total
Yes	40	160	200
No	60	740	800
Total	100	900	1,000

→ Selection Weights

100	10
100	10

→ Study Population

Exposed	Disease Yes	No	Total
Yes	40	16	56
No	60	74	134
Total	100	90	190

Target Population		Study Population
10%	Prevalence of Disease	53%
20%	Prevalence of Risk Factor	29%
2.67	Relative Risk	1.60
3.08	Odds Ratio	3.08
125/1,000	Attributable Risk	260/1,000

Table 6.8 Effects of losses to follow-up that differ by both exposure and outcome status on estimates of disease occurrence and association

Target Population

Exposed	Disease Yes	No	Total
Yes	40	160	200
No	60	740	800
Total	100	900	1,000

→ Selection Weights

50	100
100	100

→ Study Population

Exposed	Disease Yes	No	Total
Yes	20	160	180
No	60	740	800
Total	80	900	980

Target Population		Study Population
10%	Prevalence of Disease	8%
20%	Prevalence of Risk Factor	18%
2.67	Relative Risk	1.48
3.08	Odds Ratio	1.54
125/1,000	Attributable Risk	36/1,000

determine whether a particular record relates to a cohort member involves some form of record linkage, determining whether the identifying data in the study file of a cohort member corresponds with the identifying data on the medical or other record of endpoint information. In the past, much of this linkage used to be done manually. Increasingly some form of computerized record linkage is performed.

Although such linkages are easier if both sets of records contain the same (national) identifying number, computerized record linkage can still be extremely efficient and less costly than individual-based follow-up. If record linkage is planned to determine endpoints in a cohort study, great care should be taken at the time of recruitment to collect sufficient identifying information for record linkage purposes; this includes full name, full date of birth, place of birth, mother's maiden name, social security number, other identifying number (if available), and current address. Further, the name and address of friends or relatives of the cohort member should also be collected, to facilitate tracing an individual if other means of tracing them have failed or if record linkage to another data source has resulted in an uncertain linkage.

In many countries, in addition to disease registries, such as cancer registries, there are other data sources that have been developed to facilitate record linkage for cohort studies and large-scale trials. These include the National Health Service Central Register in the UK, the Canadian National Mortality Database, the National Death Index in the USA, and similar national registers in the Scandinavian countries. Relatively new in this context are the population-wide databases of genetic data, like the biobank already established in Estonia (Estonian Genome Center 2011). Record linkage using these national databases overcomes many of the issues regarding confidentiality of data, as confidentiality procedures are readily available for such systems. In Canada, what is returned to the investigator is generally anonymous (i.e., stripped of personal identifiers), unless the subjects have signed a prior consent form that specifically permitted record linkage. This was the case, for example, in a cohort study that was linked to a large multicenter trial of breast screening (Howe et al. 1991).

6.5 Ethical Issues

It is now generally accepted that studies on humans should be carried out with informed consent. This principle, originally developed in relation to controlled clinical trials, has generally now been extended to observational epidemiology studies, including cohort studies.

In the past, if a cohort was recruited that involved the subject's participation in providing data, their agreement to supply the data (e.g., respond to a questionnaire) was generally regarded as implied consent. However, now, in addition to providing information on questionnaires, for many cohorts, biological specimens (e.g., blood, buccal cells) are requested, and then it becomes mandatory that the respondent provides consent for the future use of such specimens for research purposes. However, at the time the specimens are provided, it is impossible to know the precise use the investigators may wish to apply to this material. An example relates to the fact that the majority of participants in the sub-cohorts of the European Prospective Investigation of Diet and Cancer (EPIC; Riboli and Kaaks 1997) provided blood specimens in the early 1990s; a few without signing a consent form, the majority did so. However, now that genetic studies are commonplace on such specimens, it has become apparent that some of the consent forms did not specifically mention genetic analyses as potential research usages. This has led to difficulties in obtaining approval for such sub-studies from human experimentation committees, some of which wanted new consent forms to be signed, specific to the genetically associated sub-study planned. Obtaining new consent, however, will become increasingly difficult as time goes on, and a number of subjects with the endpoint of interest may have died. In the United States, potential restrictions upon studies such as these have caused difficulties. In Europe, especially Scandinavia, there has been a more relaxed view of the ethical acceptability of studies on stored specimens, many such collections having been originally made without a formal informed consent process, but for which studies conducted, often involving record linkage to national registries, have been deemed to be ethically acceptable as full preservation of confidentiality can readily be achieved.

The issue as to whether respondents whose stored specimens have been tested should be informed of the results of such tests is also controversial. The European view tends to be that as the testing is being conducted as part of research, it may be impossible to interpret the results of tests for individuals, until this particular research track reaches agreed conclusions. Thus, it is not necessary, indeed possibly unethical, to inform the respondent of the results. Some consent forms specifically state this as a policy. In the United States, however, the opposite viewpoint tends to hold, it being regarded as ethically inappropriate for investigators to take a decision on whether or not a subject receives information on themselves. The difficulty with a universal application of such a principle is that for some, the test results may come too late for any possibility of benefit, but, especially in the case of genetic-related information, this may not preclude the test result having implications for the relatives of the subject, and such knowledge is not always a blessing. However, all would agree that if a test reveals information of potential benefit to a subject, they should be informed.

The question of consent for historical cohort studies in general does not arise, though again there may be issues on informing surviving subjects of the findings of the research. In general, as the research is unlikely to harm the individuals and providing confidentiality is maintained, human experimentation committees will approve such studies.

One further ethical issue has already been mentioned in Sect. 6.4.3, and that relates to the use of record linkage in obtaining outcome data. In general, providing full confidentiality is maintained, and this should not cause difficulties in obtaining approval from human experimentation committees. For further discussions of ethical aspects, we refer to (chapter ▸Ethical Aspects of Epidemiological Research of this handbook).

6.6 Conclusions

Cohort studies are a critical method for evaluating causality in epidemiology and may also be used in evaluating screening (see chapter ▸Screening of this handbook). There are, however, several needs if they are to be valid. You need skilled investigators familiar with the peculiarities of the planning and the conduct of cohort studies, a sensible source for cohort recruitment, evaluable hypotheses to consider, a validated questionnaire for use at enrolment, unbiased mechanisms to administer the questionnaire as well as for follow-up, quality-controlled procedures to collect biological material if relevant for the question under research, facilities for data entry, and of course the expertise as well as the facilities for analysis and interpretation.

Cohort studies are often rated at a higher level than case-control studies, largely because the latter are susceptible to recall bias. However, both are usually regarded as "level II" evidence (level I are randomized controlled trials), and there are potential deficiencies in cohort studies that may be less intrusive than in case-control studies, especially a greater propensity for measurement error. Both, however, continue to have an important role in disease epidemiology.

References

Blair A, Stewart PA (1992) Do quantitative exposure assessments improve risk estimates in occupational studies of cancer? Am J Ind Med 21:53–63

Benichou J (1998) Absolute risk. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics. Wiley, New York

Breslow NE, Day NE (1987) Statistical methods in cancer research. Volume II: The design and analysis of cohort studies. IARC scientific publications no. 82. International Agency for Research on Cancer, Lyon

Case RA, Hosker ME, McDonald DB, Pearson JT (1954) Tumours of the urinary bladder in workmen engaged in the manufacture and use of certain dyestuff intermediates in the British chemical industry. Br J Ind Med 11:75–104

Cho E, Smith-Warner SA, Spiegelman D, Beeson WL, van den Brandt PA, Colditz GA, Folsom AR, Fraser GE, Freudenheim JL, Giovannucci E, Goldbohm RA, Graham S, Miller AB, Pietinen P, Potter JD, Rohan TE, Terry P, Toniolo P, Virtanen MJ, Willett WC, Wolk A, Wu K, Yaun S-S, Zeleniuch-Jacquotte A, Hunter DJ (2004) Dairy foods, calcium, and colorectal cancer: a pooled analysis of 10 cohort studies. J Natl Cancer Inst 96:1015–1022

Coleman M, Douglas A, Hermon C, Peto J (1986) Cohort study analysis with a FORTRAN computer program. Int J Epidemiol 15:134–137

Day NE, Ferrari P (2002) Some methodological issues in nutritional epidemiology. In: Riboli E, Lambert A (eds) Nutrition and lifestyle: opportunities for cancer prevention. IARC scientific publications no. 156. International Agency for Research on Cancer, Lyon, pp 5–10

Doll R, Hill AB (1954) The mortality of doctors in relation to their smoking habits: a preliminary report. Br Med J 1:1451–1455

Doll R, Peto R (1976) Mortality in relation to smoking: 20 years' observations on male British doctors. Br Med J 2:1525–1536

Doll R, Peto R, Hall E, Wheatley K, Gray R (1994a) Mortality in relation to consumption of alcohol: 13 years' observations on male British doctors. Br Med J 309:911–918

Doll R, Peto R, Wheatley K, Gray R, Sutherland I (1994b) Mortality in relation to smoking: 40 years' observations on male British doctors. Br Med J 309:901–911

Eigenbrodt ML, Mosley TH, Hutchinson RG, Watson RL, Chambless LE, Szklo M (2001) Alcohol consumption with age: a cross-sectional and longitudinal study of the Atherosclerosis Risk in Communities (ARIC) study, 1987–1995. Am J Epidemiol 153:1102–1111

Estonian Genome Center (2011) Estonian Genome Center 2001–2011. University of Tartu. http://www.geenivaramu.ee/documents/estoniangenomecenter.pdf. Accessed 3 Mar 2012

Flanders WD, Lally CA, Zhu BP, Henley SJ, Thun MJ (2003) Lung cancer mortality in relation to age, duration of smoking, and daily cigarette consumption: results from Cancer Prevention Study II. Cancer Res 63:6556–6562

Goldberg M, Luce D (2001) Selection effects in epidemiological cohorts: nature, causes and consequences. Rev Epidemiol Sante Publique 49:477–492

Greenland S, Robins JM (1988) Conceptual problems in the definition and interpretation of attributable fractions. Am J Epidemiol 128:1185–1197

Howe GR, Friedenreich CM, Jain M (1983) A follow-up of a ten-percent sample of the Canadian labor force. I. Cancer mortality in males, 1965–73. J Natl Cancer Inst 70:37–44

Howe GR, Friedenreich CM, Jain M, Miller AB (1991) A cohort study of fat intake and risk of breast cancer. J Natl Cancer Inst 83:336–340

IARC Working Group (2003) IARC handbooks on cancer prevention, vol 8: Fruits and vegetables. International Agency for Research on Cancer, Lyon

Kleinbaum DG, Kupper LL, Morgenstern H (1982) Epidemiologic research: principles and quantitative methods. Wiley, New York

Koushik A, Hunter DJ, Spiegelman D, Anderson KE, Buring JE, Freudenheim JL, Goldbohm RA, Hankinson SE, Larsson SC, Leitzmann M, Marshall JR, McCullough ML, Miller AB, Rodriguez C, Rohan TE, Ross JA, Schatzkin A, Schouten LJ, Willett WC, Wolk A, Zhang SM, Smith-Warner SA (2006) Intake of the major carotenoids and the risk of epithelial ovarian cancer in a pooled analysis of 10 cohort studies. Int J Cancer 119:2148–2154

Li CY, Sung FC (1999) A review of the healthy worker effect in occupational epidemiology. Occup Med 49:225–229

Liddell FD, McDonald AD, McDonald JC (1997) The 1891–1920 birth cohort of Quebec chrysotile miners and millers: development from 1904 and mortality to 1992. Ann Occup Hyg 41:13–36

McCray E (1986) Occupational risk of the acquired immunodeficiency syndrome among health care workers. N Engl J Med 314:1127–1132

Miller AB, To T, Agnew DA, Wall C, Green LM (1996) Leukemia following occupational exposure to 60-Hz electric and magnetic fields among Ontario electric utility workers. Am J Epidemiol 144:150–160

Miller AB, Altenburg H-P, Bueno de Mesquita B, Boshuizen HC, Agudo A, Berrino F, Gram IT, Janson L, Linseisen J, Overvad K, Rasmuson T, Vineis P, Lukanova A, Allen N, Amiano P, Barricarte A, Berglund G, Boeing H, Clavel-Chapelon F, Day NE, Hallmans G, Lund E, Martinez C, Navarro C, Palli D, Panico S, Peeters PH, Quiros JR, Tjonneland A, Tumino R, Trichopoulou A, Trichopoulos D, Slimani N, Riboli E (2004) Fruits and vegetables and lung cancer: findings from the European Prospective Investigation into Cancer and Nutrition. Int J Cancer 108 269–276

Morgan RW, Kelsh MA, Zhao K, Exuzides KA, Heringer S, Negrete W (2000) Radiofrequency exposure and mortality from cancer of the brain and lymphatic/hematopoietic systems. Epidemiology 11:118–127

Morris JN, Heady JA, Raffle PA, Roberts CG, Parks JW (1953a) Coronary heart disease and physical activity of work. Lancet 265:1053–1057

Morris JN, Heady JA, Raffle PA, Roberts CG, Parks JW (1953b) Coronary heart disease and physical activity of work. Lancet 265:1111–1120

National Institutes of Health (2004) Request for information: design and implementation of a large-scale prospective cohort study of genetic and environmental influences on common diseases. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-04-041.html Accessed 1 Mar 2012

Oomen CM, Ocke MC, Feskens EJ, van Erp-Baart MA, Kok FJ, Kromhout D (2001) Association between trans fatty acid intake and 10-year risk of coronary heart disease in the Zutphen Elderly Study: a prospective population-based study. Lancet 357:746–751

Orencia AJ, Petty GW, Khandheria BK, Annegers JF, Ballard DJ, Sicks JD, O'Fallon WM, Whisnant JP (1995) Risk of stroke with mitral valve prolapse in population-based cohort study. Stroke 26:7–13

Preston DL, Lubin JH, Pierce DA, McConney ME (1993) Epicure user's guide. HiroSoft International Corp, Seattle

Riboli E, Kaaks R (1997) The EPIC project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. Int J Epidemiol 26(Suppl 1):S6–S14

Rinsky RA, Smith AB, Hornung R, Filloon TG, Young RJ, Okun AH, Landrigan PJ (1987) Benzene and leukemia. An epidemiologic risk assessment. New Engl J Med 316:1044–1050

Rittgen W, Becker N (2000) SMR analysis of historical follow-up studies with missing death certificates. Biometrics 56:1164–1169

Rothman KJ, Greenland S (1998) Modern epidemiology, 2nd edn. Lippincott Williams & Wilkins, Philadelphia

Sahai H, Khurshid A (1996) Statistics in epidemiology. Methods, techniques, and applications. CRC, Boca Raton/New York/London/Tokyo

Selikoff IJ, Churg J, Hammond EC (1965) The occurrence of asbestosis among insulation workers in U.S.A. Ann N Y Acad Sci 132:139–155

Silcocks P (1994) Estimating confidence limits on a standardised mortality ratio when the expected number is not error free. J Epidemiol Community Health 48:313–317

Smith-Warner SA, Spiegelman D, Yaun SS, Albanes D, Beeson WL, van den Brandt PA, Feskanich D, Folsom AR, Fraser GE, Freudenheim JL, Giovannucci E, Goldbohm RA, Graham S, Kushi LH, Miller AB, Pietinen P, Rohan TE, Speizer FE, Willett WC, Hunter DJ (2003) Fruits, vegetables and lung cancer: a pooled analysis of cohort studies. Int J Cancer 107:1001–1011

Smith-Warner SA, Spiegelman D, Ritz J, Albanes D, Beeson WL, Bernstein L, Berrino F, van den Brandt PA, Buring JE, Cho E, Colditz GA, Folsom AR, Freudenheim JL, Giovannucci E, Goldbohm RA, Graham S, Harnack L, Horn-Ross PL, Krogh V, Leitzmann MF, McCullough ML, Miller AB, Rodriguez C, Rohan TE, Schatzkin A, Shore R, Virtanen M, Willett WC, Wolk A, Zeleniuch-Jacquotte A, Zhang SM, Hunter DJ (2006) Methods for pooling results of epidemiologic studies. The pooling project of prospective studies of diet and cancer. Am J Epidemiol 163:1053–1064

Snow J (1855) On the mode of communication of cholera. Churchill, London

Sutherland J (2002) EXTRACTS from appendix (A) to the Report of the General Board of Health on the Epidemic Cholera of 1848 & 1849. Int J Epidemiol 31:900–907

StataCorp (2001) Stata statistical software: release 7.0. Stata Corporation, College Station

Terris M (ed) (1964) Goldberger on pellagra. Louisiana State University Press, Baton Rouge

Theriault G, Goldberg M, Miller AB, Armstrong B, Guenel P, Deadman J, Imbernon E, To T, Chevalier A, Cyr D (1994) Cancer risks associated with occupational exposure to magnetic fields among electric utility workers in Ontario and Quebec, Canada, and France: 1970–1989. Am J Epidemiol 139:550–572

Tomatis L, Aitio A, Day NE, Heseltine E, Kalder J, Miller AB, Parkin DM, Riboli E (eds) (1990) Cancer: causes, occurrence and control. IARC scientific publications No. 100. International Agency for Research on Cancer, Lyon

Ulvestad B, Kjaerheim K, Martinsen JI, Mowe G, Andersen A (2004) Cancer incidence among members of the Norwegian trade union of insulation workers. J Occup Environ Med 46:84–89

Villeneuve PJ, Agnew DA, Corey PN, Miller AB (1998) Alternate indices of electric and magnetic field exposures among Ontario electrical utility workers. Bioelectromagnetics 19:140–151

Wada S, Miyanishi M, Nishimoto Y, Kambe S, Miller RW (1968) Mustard gas as a cause of respiratory neoplasia in man. Lancet 1:1161–1163

Willett WC, Hunter DJ, Stampfer MJ, Colditz G, Manson JE, Spiegelman D, Rosner B, Hennekens CH, Speizer FE (1992) Dietary fat and fiber in relation to risk of breast cancer. An 8-year follow-up. J Am Med Assoc 268:2037–2044

Wolk A, Bergstrom R, Hunter D, Willett W, Ljung H, Holmberg L, Bergkvist L, Bruce A, Adami HO (1998) A prospective study of association of monounsaturated fat and other types of fat with risk of breast cancer. Arch Intern Med 158:41–45

Case-Control Studies

7

Norman E. Breslow

Contents

N.E. Breslow
Department of Biostatistics, University of Washington, Seattle, WA, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_7,

7.1 Introduction

7.1.1 A Brief History

The case-control study examines the association between disease and potential risk factors by taking separate samples of diseased cases and of controls at risk of developing disease. Information may be collected for both cases and controls on genetic, social, behavioral, environmental, or other determinants of disease risk. The basic study design has a long history, extending back at least to Guy's 1843 comparison of the occupations of men with pulmonary consumption to the occupations of men having other diseases (Lilienfeld and Lilienfeld 1979). Beginning in the 1920s, it was used to link cancer to environmental and hormonal exposures. Broders (1920) discovered an association between pipe smoking and lip cancer; Lane-Claypon (1926), who selected matched hospital controls, investigated the relationship between reproductive experience and female breast cancer; and Lombard and Doering (1928) related pipe smoking to oral cancer. The publication in 1950 of three reports on the association between cigarette smoking and lung cancer generated enormous interest in case-control methodology as well as bitter criticism (Levin et al. 1950; Wynder and Graham 1950; Doll and Hill 1950). The landmark study of Doll and Hill (1950, 1952), in particular, inspired future generations of epidemiologists to use this methodology. It remains to this day a model for the design and conduct of case-control studies, with excellent suggestions on how to reduce or eliminate selection, interview, and recall bias.

From the mid-1950s to the mid-1970s the number of case-control studies published in selected medical journals increased four- to sevenfold (Cole 1979). Aird et al. (1953) discovered the association between gastric cancer and the ABO blood groups. The impact of hormonal factors on cancers of female organs was brought to light, starting with confirmation of the association between late first pregnancy and breast cancer (MacMahon et al. 1970). Herbst et al. (1971) investigated an unusual outbreak of vaginal adenocarcinoma in young women, finding that mothers of seven of eight cases had exposed their daughters *in utero* to the fertility drug diethylstilbestrol (DES). None of 32 control mothers had a history of estrogen use during pregnancy. Treatment of menopausal women with exogenous estrogens similarly increased the risk of endometrial cancer (Ziel and Finkle 1975; Smith et al. 1975). Powerful joint effects of alcohol and tobacco consumption on esophageal cancer were demonstrated (Tuyns et al. 1977), as was the strong association between liver cancer and Hepatitis B carrier status (Prince et al. 1975). These successes encouraged more investigators to adopt the case-control study as the method of choice for the study of rare chronic diseases, particularly cancer. A survey by Correa et al. (1994) identified 223 population-based case-control studies published in the world literature in 1992. Subsequent discoveries obtained using case-control methodology have included the role of salted fish in the etiology of nasopharyngeal carcinoma in Chinese populations (Armstrong et al. 1983; Yu et al. 1986), the hazards of prone sleeping position for sudden infant death syndrome (SIDS) (Fleming et al. 1990), and the relationship between use of intrauterine devices (IUDs) and tubal infertility (Daling et al. 1985).

This plethora of case-control studies, stimulated by their relatively low cost and short duration, also had its drawbacks. Not all investigators were as careful as Doll and Hill in following a protocol for selection of cases and controls, in conducting the study to mitigate against bias, and in thoughtfully analyzing the collected data. Nor did they have the good fortune to study associations as strong as that between lung cancer and cigarette smoking. The increasing availability of high-speed computers made it possible to collect more and more data and to look for all manner of associations with putative risk factors. Investigators eager for research funding were sometimes too quick to publish their findings and draw media attention to them. The inevitable result was an increasingly negative reaction on the part of the public, and from segments of the scientific community, to the false alarms and contradictory results (Taubes 1995). One goal of this chapter, and of others in this handbook, is to describe basic scientific principles whose application should help to improve public confidence in published findings of epidemiological studies.

7.1.2 Early Methodological Developments

> The sophisticated use and understanding of case-control studies is the most outstanding methodological development of modern epidemiology.
>
> (Rothman 1986, p. 62)

The initial interpretation of the case-control study was the comparison of exposure histories for a group of diseased cases with those for non-diseased controls. Typical analyses involved two group comparisons of exposure distributions using chi-squared and t-tests. The critics argued that such comparisons provided no information about the quantities of true epidemiological interest, namely, the disease rates. Cornfield (1951) corrected this misconception by demonstrating that the exposure odds ratio for cases vs. controls was equal to the disease odds ratio for exposed vs. non-exposed. With $D = 1$ indicating disease, $D = 0$ disease-free, and $X = 1/0$ likewise denoting exposed or non-exposed, he showed using Bayes theorem that

$$\frac{\Pr(D = 1|X = 1)\Pr(D = 0|X = 0)}{\Pr(D = 0|X = 1)\Pr(D = 1|X = 0)} = \frac{\Pr(X = 1|D = 1)\Pr(X = 0|D = 0)}{\Pr(X = 0|D = 1)\Pr(X = 1|D = 0)} \tag{7.1}$$

and noted that the disease odds ratio approximated the *relative risk* $\Pr(D = 1|X = 1)/\Pr(D = 1|X = 0)$ provided the disease was rare. He also pointed out that, if the overall disease risk was known from other data sources, this could be combined with the relative risk to estimate *absolute* disease risks for exposed and non-exposed, respectively.

Disease risk as considered by Cornfield (1951) was *prevalence*, the probability that a member of the population was ill at a given point in time. For studies of disease etiology, however, it is preferable to work with disease *incidence*, the probability of developing disease during the study period among subjects who are free of disease initially. Otherwise, one confuses the effect of exposure on causation of disease with its effect on the case fatality rate (Neyman 1955). Controls for a study of the cumulative risk of developing disease during a given period would be persons who were free of disease during the entire period. Although it laid the foundation for what was to follow, this conceptualization of the case-control study in terms of cumulative disease risk was awkward for two reasons. First, as the study interval lengthened the risk of disease increased for both exposed and non-exposed. The relative risk for a common disease could approach one. Even if it did not, it was undesirable to have the basic effect measure so dependent on study duration, which varies between studies. Second, for a study of long duration, ensuring that the controls were disease-free throughout the study period could be problematic in practice. The modern conception of a case-control study involves sampling of controls who are disease-free at random times during the study period (Sect. 7.2.1). Exposure odds ratios are used to estimate ratios of incidence *rates* rather than ratios of risks. No rare disease assumption is needed in this case.

Mantel and Haenszel (1959) clarified the status of the case-control (or retrospective) study in comparison with the cohort (forward or prospective) study in one of the most highly cited papers in the scientific literature (Breslow 1996). They stated emphatically:

> A primary goal is to reach the same conclusions in a retrospective study as would have been obtained from a forward study, if one had been done.
>
> (Mantel and Haenszel 1959, p. 722)

This insight underlies the modern conception of the case-control study as involving *sampling*, on the basis of outcome, from an ongoing real or imagined cohort study that has been designed to provide the best possible answer to the basic question. Mantel and Haenszel introduced a new test and a simple, highly efficient estimator for the relative risk after stratification on control factors. Their methods required the epidemiologist to carefully examine the tabular data and thus to identify strata where there was a lack of information or where there were discrepancies between summary and stratum specific relative risks. They remain valuable today as an adjunct to more elaborate model fitting.

By the end of the 1950s, the case-control study was firmly established as the method of choice for the chronic disease epidemiologist, certainly when the budget was limited. The role of statisticians in bringing the study design to this place of scientific respectability was widely acknowledged (Cole 1979; Armenian and Lilienfeld 1994). Further methodological advances were made during the next two decades, particularly in statistical modeling of case-control data. The development of the proportional hazards regression model for life table data (Sheehe 1962; Cox 1972) provided a sound mathematical basis for methods long used by epidemiologists and led to refinements and extensions of those methods (Breslow et al. 1983). The nested case-control study, originally conceived as a method to reduce the computational burden of fitting Cox's model to data from large cohorts (Liddell et al. 1977), was recognized as an efficient epidemiological design for the collection of expensive explanatory data (Langholz and Goldstein 1996). It now serves as a paradigm for all case-control studies. Many of the methodological developments were described in texts by Breslow and Day (1980) and Schlesselman (1982) that led to further appreciation and use of the case-control study.

7.1.3 Chapter Outline

The remainder of this chapter discusses the modern conceptualization of the case-control study, largely from a statistical perspective. Matching of controls to cases at the design stage is viewed as a technique to be used in carefully limited contexts to increase the statistical efficiency of a highly stratified analysis. The implications of these theoretical developments for the practical selection of cases and controls are explored. Major pitfalls include the unique susceptibility of the case-control study to selection bias and, especially when exposures are assessed by interview, to measurement error. The design of any particular study usually involves trade-offs between potential biases arising from these sources. Following established principles of sound statistical science, including the use of an appropriate protocol for subject selection and exposure assessment, can help reduce the variability in study results that has contributed to the low esteem accorded risk factor epidemiology in some scientific circles (Breslow 2003).

7.2 Conceptual Foundations

7.2.1 Sampling from a Real or Fictitious Cohort

The Mantel and Haenszel (1959) goal, of reaching the same conclusions from a case-control study as from a cohort study if one had been done, provides the key to understanding of case-control methodology. Rather than start the planning process by thinking about how to conduct a case-control study, it often is helpful to first plan the ideal cohort study that would be conducted to investigate the same hypothesis if unlimited resources were available. Planning would include cohort identification, definition of the times of entry into and exit from the cohort, ascertainment of the disease endpoint, measurement of the exposure histories, consideration of potential confounders, and methods of statistical analysis. The corresponding case-control study would then be viewed as the random sampling of subjects from this idealized cohort to achieve, so far as possible, the stated goal.

Cohort Definition In concept, the underlying cohort for a case-control study consists of all subjects who, had they experienced the disease endpoint at a specific time, would have been ascertained as a case at that time. When case-control sampling is carried out in the context of an actual cohort study, to select individuals for genotyping or other expensive measurements, for example, the cohort is completely enumerated by and known to the investigator. More typically, however,

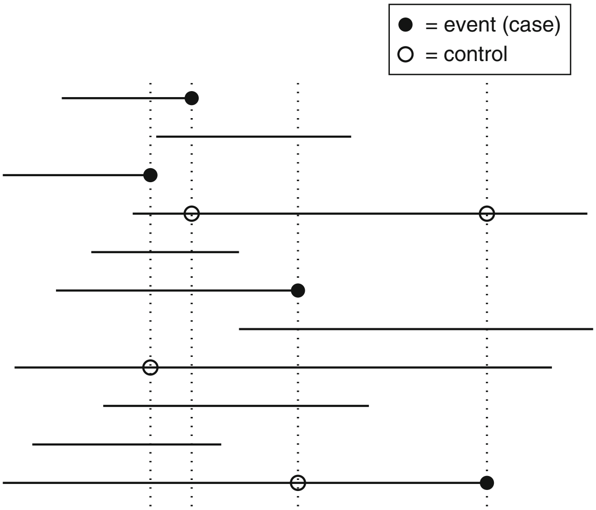


Fig. 7.1 Schematic of a (nested) case-control study

the underlying cohort is not fully identified and is effectively defined by the method of case ascertainment. When cases are ascertained from a particular hospital, for example, one considers the cohort to consist of all subjects who, had they developed the disease in question, would have been diagnosed in that hospital.

Figure 7.1 illustrates the basic idea of case-control sampling. Each of the 11 horizontal lines represents time on study for a member of the cohort. Subjects enter follow-up at the left-hand endpoint and exit at the right. They are considered to be *at risk* of becoming a case throughout this period. It is even possible, though not shown here, that a subject could enter the cohort, leave for a while and then return. Four of the 11 subjects are cases. Their follow-up ends at diagnosis since they are no longer at risk of becoming an incident case thereafter. The vertical dotted lines, plotted at each of the times that a case occurs, intersect the trajectories of those who are at risk at that time, that is, the trajectories of subjects in the corresponding *risk set*.

Nested Case-Control Sampling When the cohort study is a real one, so that times of entry and exit are known for all members, the investigator may completely enumerate each risk set. A nested case-control study is then possible in which controls are selected by finite population random sampling, without replacement, from non-cases in the risk set. The usual assumption is that the sampling of controls from each risk set is completely independent of sampling from all other risk sets. Two consequences are that a subject sampled as a control at one point in time may later become a case and that the same subject may be sampled more than once as a control. Figure 7.1, which depicts the situation where exactly one control is sampled from each risk set, illustrates each of these possibilities.

Robins et al. (1986) describe other sampling schemes, and corresponding methods of analysis, for nested case-control studies.

Density Sampling These ideas also may be applied, at least in principle, to the more typical situation in which the cohort is not completely enumerated. An essential assumption, which in fact well approximates the design of many studies, is that the cohort is sampled throughout the study period. More specifically, controls are selected at any given time at a rate proportional to the disease incidence rate at that time (Sheehe 1962). Miettinen (1976) termed this *incidence density sampling*. A second assumption is that each subject at risk at a given time has the same probability of being sampled as a control. This implies that, from the standpoint of an individual, the likelihood of being included in the study as a control increases with increasing *time on study*. If the disease incidence rate is constant, someone who is a member of the cohort for twice as long as someone else has twice the chance of being selected as a control. In the statistical literature, this is known as *length biased* sampling. One important consequence, under the assumption of constant disease incidence, is that the number of controls sampled is proportional to the *total time at risk*.

7.2.2 Incidence Rate Ratios are Estimable from Odds Ratios

We consider here the simplest situation in which the disease incidence rate is constant and there are two groups of subjects, exposed and non-exposed, that are homogeneous apart from exposure. Confounding is therefore not an issue. Denote by A the total number of incident cases ascertained from the cohort during the study period (t_0, t_1) and suppose that A_0 are determined to be non-exposed whereas $A_1 = A - A_0$ are exposed. Similarly, denote by $T = T_0 + T_1$ the total person-time on study, decomposed into its non-exposed (T_0) and exposed (T_1) components. While the numbers of cases A_0 and A_1 are known to the investigator, T_0 and T_1 may not be unless the underlying cohort is a real one. Instead, the case-control study provides information on how many of the total $M = M_0 + M_1$ of controls are non-exposed (M_0) and how many are exposed (M_1). Denoting by M_1/M_0 the observed odds of exposure for controls and likewise by A_1/A_0 the observed odds of exposure for cases, the corresponding exposure odds ratio is $(A_1M_0)/(A_0M_1)$.

Let $\pi\tau$ denote the probability that a subject who contributes τ person-years of follow-up is sampled as a control. With $T = \sum_{i=1}^{N} \tau_i$ denoting the sum of the times on study for N cohort members, that is, the total time at risk, the expected number of controls is $E(M) = \pi T$. In practice, π is often selected by the investigator to yield a fixed number of controls, at least as a target value. Its actual value remains unknown unless information is available about T. Nonetheless, provided π is constant for all subjects, both exposed and non-exposed, $E(M_0) = \pi T_0$ and $E(M_1) = \pi T_1$. Hence, the control ratio M_0/M_1 estimates the corresponding ratio T_0/T_1 of person-time. Since the exposure specific incidence rates are estimated by $\hat{\lambda}_0 = A_0/T_0$ and $\hat{\lambda}_1 = A_1/T_1$, it follows (see Rothman and Greenland 1998, Chap. 10) that the rate ratio may be estimated by the exposure odds ratio:

$$\frac{\hat{\lambda}_1}{\hat{\lambda}_0} = \frac{A_1 T_0}{A_0 T_1} \approx \frac{A_1 M_0}{A_0 M_1} \ . \tag{7.2}$$

See Sect. 7.3.1 for a numerical example.

7.2.3 Time-Dependent Rates and Exposures

Section 7.2.2 assumes that the parameter of interest is the ratio of instantaneous incidence rates, each assumed constant in time, for exposed and non-exposed subjects. A more general conceptualization takes the interest parameter to be the ratio $\psi \equiv \lambda_1(t)/\lambda_0(t)$ of instantaneous rates where the ratio, but not necessarily the underlying rates, is assumed constant in t. Let $N(t)$ denote the total number of subjects at risk at time t in the underlying cohort, of which a proportion $p_1(t)$ are exposed and $p_0(t)$ are non-exposed. These proportions could vary with time either because the exposure status for individual subjects changes or because the exposure composition of the cohort changes through entries and exits. Note that the expected number of exposed cases is given by $\int N(t)p_1(t)\lambda_1(t)\mathrm{d}t$ and similarly for the non-exposed cases. The expected number of controls sampled in the interval $(t, t + \mathrm{d}t)$ is therefore $M(t)\mathrm{d}t$ where $M(t) = N(t)[p_0(t)\lambda_0(t) + p_1(t)\lambda_1(t)]$. It follows that the unadjusted exposure odds ratio under density sampling estimates:

$$\begin{aligned}\psi^* &= \frac{\int N(t)p_1(t)\lambda_1(t)\mathrm{d}t \int M(t)p_0(t)\mathrm{d}t}{\int N(t)p_0(t)\lambda_0(t)\mathrm{d}t \int M(t)p_1(t)\mathrm{d}t} \\ &= \psi \frac{\int N(t)p_1(t)\lambda_0(t)\mathrm{d}t \int M(t)p_0(t)\mathrm{d}t}{\int N(t)p_0(t)\lambda_0(t)\mathrm{d}t \int M(t)p_1(t)\mathrm{d}t}\end{aligned} \tag{7.3}$$

(Greenland and Thomas 1982). Thus, the exposure odds ratio estimates the incidence rate ratio, that is, $\psi^* = \psi$, provided either that the exposure proportions are constant in t or else that $\psi = 1$. Otherwise, time t acts as a *confounder* of the exposure–disease association. In this case, a time-matched analysis using standard methods for matched case-control studies (Breslow and Day 1980, Chap. 7) is needed to estimate ψ unbiasedly. The marginal (unmatched) odds ratio usually provides a slightly *conservative* estimate of this parameter.

7.2.4 Cumulative Risk Ratios and Case-Cohort Sampling

While it is generally agreed that case-control studies of chronic disease are best designed using density sampling to estimate the incidence rate ratio, alternative sampling designs may be superior for other purposes. Vaccine efficacy is usually defined as the proportional reduction, over the study period, in the number of cases among subjects who are vaccinated compared to those who are not. Equivalently, it is 1 minus the ratio of cumulative disease risks for vaccinated vs. non-vaccinated. Suppose the effect of vaccination is to render completely immune a proportion P_{I} of subjects, while the remainder of those vaccinated have the same disease incidence rates $\lambda_0(t)$ as do non-vaccinated persons (Smith et al. 1984). For simplicity, assume that all subjects, both vaccinated and non-vaccinated, are followed from a common starting time t_0 and that there is no loss to follow-up. The cumulative risk of disease by time t_1 for those not vaccinated is $P(t_0, t_1) = 1 - \exp[-\int_{t_0}^{t_1} \lambda_0(t)\mathrm{d}t]$, and the vaccine efficacy is thus

$$1 - \frac{\text{risk for vaccinated}}{\text{risk for non-vaccinated}} = 1 - \frac{P_{\mathrm{I}} \times 0 + (1 - P_{\mathrm{I}}) \times P(t_0, t_1)}{P(t_0, t_1)} = P_{\mathrm{I}}\,. \quad (7.4)$$

Here, the cumulative risk ratio, not the incidence rate ratio, is independent of study duration $t_1 - t_0$ (Rodrigues and Kirkwood 1990). Suppose now a *subcohort* of M subjects is drawn at random from the combined cohort of vaccinated and non-vaccinated subjects such that each individual has the *same* probability π of inclusion in it, *regardless* of duration of follow-up. If M_0 and M_1 denote the numbers of non-vaccinated and vaccinated in the subcohort, while A_0 and A_1 denote the numbers of disease cases diagnosed by time t_1, then vaccine efficacy is simply estimated as

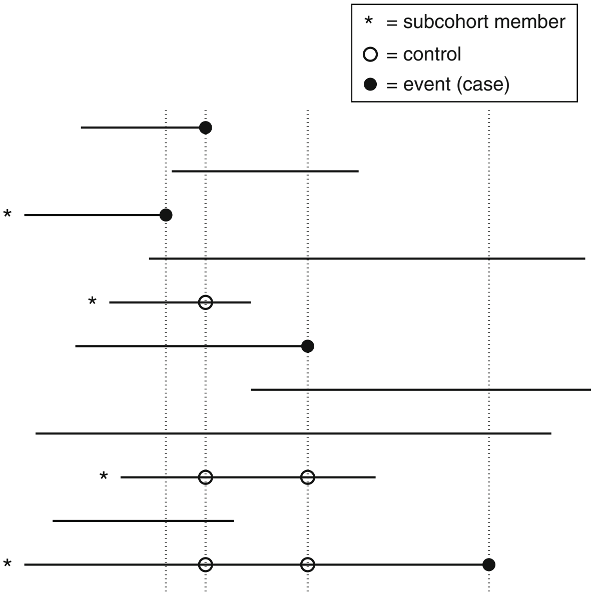
$$\widehat{P}_{\mathrm{I}} = 1 - \frac{A_1/M_1}{A_0/M_0}\,. \quad (7.5)$$

More generally, the case-cohort design (Kupper et al. 1975; Miettinen 1982; Prentice 1986) involves random sampling of a subcohort at study entry, without regard to time on study. Figure 7.2 contrasts this design with nested case-control sampling (Fig. 7.1). Incidence rate ratios may be estimated for dynamic (open) cohorts, with staggered entry and loss to follow-up as pictured, just as they are with nested case-control sampling (Prentice 1986; Lin and Ying 1993; Barlow 1994). Subcohort members under observation at the time of disease occurrence serve as the controls for each case in a time-matched analysis. Since the subcohort is a simple random sample from the full cohort, it is suitable for estimation of population genotype or exposure frequencies, whereas the controls from a nested study are not. Furthermore, the same subcohort may be used to provide controls for two or more different types of disease cases. Because of this flexibility, the case-cohort design is increasingly used for sampling from defined cohorts.

7.2.5 Estimation of Absolute Risks

The key feature of case-control sampling in the context of an actual cohort study, where the underlying cohort is completely enumerated and entry and exit times are known for all cohort members, is that the sampling probabilities for cases and

Fig. 7.2 Schematic of a case-cohort study



controls are known or can be estimated from the available data. The case-control study provides supplementary information on explanatory variables for a randomly selected group of cohort members. Analysis of the combined cohort and case-control data may be approached using standard methods for incomplete data (Little and Rubin 2002). The Horvitz and Thompson (1952) survey sampling approach is often easiest to implement. Here, the contribution to estimators or estimating equations from each subject with complete data, that is, each subject included in the case-control sample, is weighted by (an estimate of) the inverse probability of having been included. Any analysis that could have been carried out were explanatory data available for the entire cohort can also be carried out using the combined data from the cohort and the case-control sample. This principle applies to estimation of absolute as well as relative risks.

A demonstration that absolute risks can be estimated from case-control data that are supplemented with information regarding the underlying population was provided by Doll and Hill (1952). They restricted the analysis to cases and controls drawn from the Greater London area, for which the numbers of persons and the numbers of deaths due to lung cancer were known from government records for each category of age and sex. Table 7.1 shows numbers of male cases n_{1j} and controls n_{0j} aged 45–64 years at the jth of 6 levels of average cigarette consumption during the preceding 10 years ($j = 1, \ldots, 6$). Assuming that the smoking habits of the controls were representative of the habits of the general population in each age-sex category, and likewise that the habits of the cases were reasonably similar to those of persons who died of lung cancer, they were able to estimate the numbers of persons N_j and of deaths D_j at each of the 6 smoking levels. Specifically, knowing

Table 7.1 Numbers of lung cancer cases and controls in Greater London among males aged 45–64 years, by average amount smoked in preceding 10 years, with estimated death rates of lung cancer per 1,000 persons per year (Reconstructed from data of Doll and Hill (1952, p. 1278))

	Ave. daily number of cigarettes						
	0	1-4	5-14	15-24	25-49	50+	Total
Controls (n_{0j})	38	87	397	279	119	12	$n_{0+} = 932$
Cases (n_{1j})	2	19	197	171	129	21	$n_{1+} = 539$
Rates ($\hat{\lambda}_j$)	0.14	0.59	1.35	1.67	2.95	4.76[a]	1.57

[a] Doll and Hill give 4.74 for this entry

that the total male population of Greater London aged 45–64 was $N_+ = 937000$, they estimated the subpopulation (in thousands) at the jth smoking level as $\widehat{N}_j = (n_{0j}/932) \times 937$. Similarly, knowing that $D_+ = 1,474$ deaths from lung cancer occurred annually in this population, they estimated the numbers of deaths at that level by $\widehat{D}_j = (n_{1j}/539) \times 1{,}474$. Thus, the absolute rates per 1,000 persons per year at smoking level j were estimated as

$$\hat{\lambda}_j = \frac{\widehat{D}_j}{\widehat{N}_j} = \frac{n_{1j} \cdot n_{0+} \cdot D_+}{n_{0j} \cdot n_{1+} \cdot N_+} .$$

See Table 7.1 and Doll and Hill (1952), Table XII. Neutra and Drolette (1978) formally justified this commonly used procedure while Greenland (1987) provided an extension for matched case-control studies.

Langholz and Borgan (1997) developed more specialized methodology for estimation of absolute risks from nested case-control studies under the Cox (1972) model. The absolute risk of disease over the time period (t_0, t_1) for a subject with explanatory variables x who is disease-free at its start is

$$P(t_0, t_1; x) = \int_{t_0}^{t_1} S(t_0, t; x)\lambda(t; x)\mathrm{d}t \ , \tag{7.6}$$

where $S(t_0, t; x)$ denotes the probability that the subject remains on study and free of disease from t_0 to t and $\lambda(t; x)$ is the disease incidence rate. Increments in the baseline cumulative incidence rate function at each time of disease diagnosis, needed to estimate both S and λ, are obtained from the usual formula for the cohort study applied to *reduced* risk sets consisting of the case and sampled control(s). The denominator term, representing the sum of relative risks for all subjects in the risk set, is weighted by n/m where n denotes the size of the risk set and m the number of subjects, including the case, sampled from it. Benichou and Gail (1995) studied similar methodology for unmatched case-control sampling from an actual cohort when all explanatory variables are discrete. Econometricians also have developed methods for incorporation of external information on background rates into the analyses of data collected in "choice-based" sampling designs, the social science analog of case-control studies (Hsieh et al. 1985).

7.3 Matching and Stratification

> While the logical absurdity of attempting to measure an effect for a factor controlled by matching must be obvious, it is surprising how often investigators must be restrained from attempting this.
>
> (Mantel and Haenszel 1959, p. 729)

Investigators planning case-control studies used to consider matching of individual controls to cases as a means of making the two groups as comparable as possible, thereby increasing the perceived validity of study results. It is now recognized that such matching, or stratified sampling of controls to make them more like the cases – known as frequency matching, has a much more limited and specific role. This is to improve the efficiency of rate ratio estimators (exposure odds ratios) that are statistically *adjusted* to account for possible confounding effects. Inappropriate matching may have the unintended effect of compromising design efficiency or even of rendering the results completely uninterpretable. Furthermore, since the sampling design must always be considered, matching usually complicates the statistical analysis.

7.3.1 Consequences of Matching

The goal of matching in case-control studies is to balance the numbers of cases and controls within strata that will be used for statistical adjustment purposes. If the factor(s) used for stratification are associated with exposure, the matched control sample will generally have an exposure distribution more like that of the cases than would an unmatched control sample.

Some interesting and important consequences of matching are illustrated by the fictitious data shown in Table 7.2, which is adapted from Table 10.5 of Rothman and Greenland (1998). In the underlying cohort, the disease rates for exposed and non-exposed are identical for males and females. Consequently, there is no effect modification nor confounding by sex, and the crude (marginal) rate ratio equals the sex-specific ratios. The frequency matching of controls to cases by sex, however, has *induced* apparent confounding in the case-control data. The sex-specific rate ratios are correctly estimated by the sex-specific odds ratios, in accordance with Eq. 7.2, but they are substantially under-estimated by the crude exposure odds ratio. An analysis that accounts for the matching is essential to correctly estimate the interest parameter.

Table 7.2 Distribution of cases and person-years of observation in a fictitious cohort study and expected distribution of cases and frequency matched controls (Adapted from Table 10.5 of Rothman and Greenland (1998))

A. Results for underlying cohort study				
	Males		Females	
	Exposed	Non-exposed	Exposed	Non-exposed
Diseased	450	10	50	90
Person-years	90,000	10000	10000	90,000
Rate ($\times 10^3$)	5.0	1.0	5.0	1.0
Rate ratio	$\psi_{\mathrm{E}} = 5$		$\psi_{\mathrm{E}} = 5$	
	Crude rate ratio = $\frac{(450+50)/100{,}000}{(10+90)/100{,}000} = 5$			
B. Expected results for the case-control study				
	Males		Females	
	Exposed	Non-exposed	Exposed	Non-exposed
Cases	450	10	50	90
Controls	414	46	14	126
Odds ratio	$\widehat{\psi}_{\mathrm{E}} = 5.0$		$\widehat{\psi}_{\mathrm{E}} = 5.0$	
	Expected crude odds ratio $\approx \widehat{\psi}_{\mathrm{E}} = \frac{(450+50)\times(46+126)}{(10+90)\times(414+14)} = 2$			

7.3.2 Efficiency of Matching

The advantages of a frequency-matched sample become evident when one considers extreme situations. In the study of esophageal cancer of Tuyns et al. (1977), for example, 775 controls were sampled at random from electoral rolls for comparison with the 200 cases. Not surprisingly, the lowest age stratum contained only a single case and 115 controls. Since they contributed very little to the age-stratified odds ratio, the time spent interviewing the 115 youngest controls was largely wasted. When the potential for imbalance is less extreme, however, the advantages of matching are not so clear. Some insight is provided by considering the ratio of asymptotic variances of crude and adjusted (stratified) odds ratio estimators for frequency-matched and random samples in the simplest of situations, that involving a binary exposure factor, a binary confounding factor, and a rare disease. Assuming equal numbers of cases and controls, and that the exposure rate ratio ψ_{E} is the same at both levels of the confounder, the variances are determined by five quantities: ψ_{E}; p_{E}, the population proportion exposed; p_{C}, the proportion positive for the confounder; ψ_{C}, the rate ratio for the confounder; and ψ_{CE}, the odds ratio associating confounder and exposure in the population. Table 7.3, adapted from Breslow (1982), shows ratios of variances and biases for different odds ratio estimators when $p_{\mathrm{C}} = 0.5$ and $p_{\mathrm{E}} = 0.3$. Similar results were given by Thomas and Greenland (1983) and by Smith and Day (1984).

A stratified analysis is not needed to control confounding when $\psi_{\mathrm{CE}} = 1$ or $\psi_{\mathrm{C}} = 1$. For as shown in rows 1–5, 9, and 13 of Table 7.3, the bias B_{R} of the pooled estimator using a randomly selected control sample is then zero. Columns labeled

Table 7.3 Variance ratios and biases, in percent, for estimators of ψ_E in case-control studies with matched and random samples (Adapted from Breslow (1982), Table 2)

		$\psi_E = 1$				$\psi_E = 2$				$\psi_E = 5$				$\psi_E = 10$			
ψ_{CE}	ψ_C	$\frac{V_M}{V_R}$	$\frac{V_M}{V_{R^*}}$	B_R	B_M	$\frac{V_M}{V_R}$	$\frac{V_M}{V_{R^*}}$	B_R	B_M	$\frac{V_M}{V_R}$	$\frac{V_M}{V_{R^*}}$	B_R	B_M	$\frac{V_M}{V_R}$	$\frac{V_M}{V_{R^*}}$	B_R	B_M
1	1	100	100	0	0	100	100	0	0	100	100	0	0	100	100	0	0
	2	97	100	0	0	97	100	0	0	97	100	0	0	97	100	0	0
	5	88	100	0	0	87	100	0	0	87	100	0	0	88	100	0	0
	10	80	100	0	0	79	100	0	0	80	100	0	0	81	100	0	0
2	1	100	103	0	0	100	103	0	−4	100	103	0	−4	101	104	0	−6
	2	95	100	12	0	96	100	12	−4	97	101	12	−4	99	102	12	−5
	5	85	97	24	0	86	97	24	−1	88	98	24	−2	91	99	24	−3
	10	78	96	31	0	78	96	31	−1	81	97	31	−1	85	98	31	−2
5	1	100	113	0	0	99	113	0	−8	101	118	0	−18	105	122	0	−23
	2	93	106	27	0	93	106	27	−7	97	110	27	−14	103	114	27	−18
	5	82	99	58	0	83	99	58	−4	89	102	58	−8	95	105	58	−10
	10	76	95	75	0	77	96	75	−2	82	98	75	−5	88	100	75	−6
10	1	100	126	0	0	98	126	0	−14	100	134	0	−29	107	144	0	−37
	2	91	114	36	0	90	114	36	−11	96	120	36	−23	104	128	36	−28
	5	80	102	82	0	81	102	82	−6	88	107	82	−13	96	111	82	−16
	10	74	97	106	0	76	98	106	−4	82	101	106	−8	90	105	106	−9

ψ_{CE} odds ratio associating exposure and confounder
ψ_E rate ratio for exposure ψ_C rate ratio for confounder
B_R bias of pooled estimate of ψ_E, random sample, as, percent of ψ_E
B_M bias of pooled estimate of ψ_E, matched sample, as percent of ψ_E
V_M variance of the stratified estimate in the matched sample
V_R variance of the stratified estimate in the random sample
V_{R^*} variance of the pooled estimate in the random sample

$V_{\mathrm{M}}/V_{\mathrm{R}}^{*}$ show the increase in variance, that is, the loss in efficiency, if a matched control sample and stratified analysis were used instead. There is no efficiency loss through matching when $\psi_{\mathrm{CE}} = 1$ but increasing loss for estimation of large rate ratios when the correlation between confounder and exposure is high. Since the "confounder" is not a risk factor for disease ($\psi_{\mathrm{C}} = 1$), it need not be controlled in the analysis. By needlessly matching on it, the exposure distributions for cases and controls have been made more alike, thus reducing the efficiency of estimation of the exposure effect. The negative biases associated with the crude analysis of the matched data reflect the same phenomenon as the example in Table 7.2. This is a case of *overmatching*.

Stratification *is* needed to control confounding when both $\psi_{\mathrm{CE}} > 1$ and $\psi_{\mathrm{C}} > 1$. Then, as shown in rows 6–8, 10–12 and 14–16 of the table, the bias B_{R} using the unadjusted design and analysis is non-zero and becomes increasingly serious as the effect of the confounder and its correlation with the exposure increase. The efficiency of the matched design to the standard design, using in both cases the correct (stratified) analysis, may be read from columns labeled $V_{\mathrm{M}}/V_{\mathrm{R}}$. Values *under* 100% indicate greater efficiency, meaning a smaller variance, for the matched design. When the potential confounder increases disease risk but exposure does not, matching is always more efficient and its efficiency increases with the degree of confounding. Even in the most extreme situation ($\psi_{\mathrm{CE}} = \psi_{\mathrm{C}} = 10$), however, no more than 26% of efficiency is lost by failure to match. A conclusion is that confounder and disease must be strongly associated for matching to produce major gains. Matching may actually *lose* efficiency when ψ_{CE} and ψ_{E} are both large.

7.3.3 Overmatching

Overmatching refers to matching on a factor that is not a confounder of the disease–exposure association. There are three possibilities.

Factor Related Only to Exposure This is the situation just considered in Tables 7.2 and 7.3 (rows 5, 9, 13). Matching is not needed to control confounding and leads to a loss of efficiency.

Factor Related Only to Disease This has been called "the case of futility" because the matching is effectively at random with respect to exposure (Miettinen 1970). Frequency matching has no effect on efficiency, as the variance ratios $V_{\mathrm{M}}/V_{\mathrm{R}*} = 100$ when $\psi_{\mathrm{CE}} = 1$ suggest. Were one to "incorrectly" fail to account for the matching in the analysis, however, there would be efficiency loss relative to the frequency matched analysis; note the percentages below 100 in the column labeled $V_{\mathrm{M}}/V_{\mathrm{R}}$. Individual pair matching in such circumstances could cause a loss of efficiency because of the need to account for this in the analysis and the consequent reduction in degrees of freedom for estimation of the main effect. With binary exposure measurements, for example, only the discordant case-control pairs would contribute to estimation of the exposure odds ratio, and these would become fewer and fewer as the association between the matching factor and disease increased.

“Confounder” an Intermediate in the Causal Pathway The most serious type of overmatching occurs when one matches on a factor that is both affected by exposure and a cause of disease. If the effect of antihypertensive medication on the risk of myocardial infarction was being investigated, for example, yet cases and controls were matched on blood pressure measurements taken after treatment commenced, the data would be completely useless for estimation of treatment effect. Ignoring the matching in the analysis would only compound the error by driving the odds ratio even closer toward unity.

7.3.4 When to Match

In view of the drawbacks of overmatching, and the often modest efficiency gains even when statistical adjustment is indicated, one may well ask whether matching is ever justified. The administrative costs of locating matched controls, and the loss of cases from analysis if none can be found, further argue for careful consideration of matched designs. Individual case-control matching is most appealing when needed to control the effects of a confounder that is not easily measured. The paradigm is use of an identical co-twin to control for genotype (Jablon et al. 1967). Otherwise, stratification of the control sample on gender and broad categories of age to achieve rough comparability with the case distribution, provided that this can be accomplished without great cost, is likely all that is advisable. Greater attention to stratification of the control sample may be needed when the primary goal is to evaluate statistical *interaction*, or effect modification, between exposure and a covariate (Smith and Day 1984).

7.4 Selection of Subjects

The two preceding sections outline the basic ideas of sampling of subjects for a case-control study from a theoretical statistical perspective. While the theory is an important guide to practice, implementation is usually imperfect and requires some compromise to minimize the various types of bias to which case-control studies are particularly susceptible (Sect. 7.5). In this section, we consider some of the choices available to the investigator for putting the theory into action.

7.4.1 Selection of Cases

7.4.1.1 Disease Definition

Careful definition of the disease endpoint to conform to the goals of the study is critical to success. Specific cancers are reasonably well defined by primary site and histologic type. Studies of diabetes, rheumatoid arthritis, or psychiatric conditions should follow standard criteria for diagnosis established by professional societies. In the typical study of disease etiology, the investigator may choose to enhance efficiency by including only those cases of disease most likely a priori to have been caused by the particular exposure. Thus, instead of "uterine cancer," studies of hormonal risk factors would best be restricted to adenocarcinoma of the endometrium whereas those investigating sexual practices or viral etiology would focus on squamous cell cancer of the cervix. Of course, in the early stages of an investigation, demonstration that the exposure effect is *specific* to a particular disease subtype can be an important part of the evidence that the association is causal (Hill 1965; Weiss 2002). For case-control studies of the public health impact of exposure, furthermore, a broader definition of disease may be desirable.

As mentioned in Sect. 7.1.2, studies of disease etiology are best restricted to incident cases. This may not always be possible, however. Congenital anomalies are generally ascertained as those that are prevalent at birth, and consideration of possible exposure effects on fetal loss forms an important part of the interpretation. Cohort studies may be preferable for estimating the true effects of exposure on reproductive outcomes (Weinberg and Wilcox 1998).

7.4.1.2 Sources of Cases

Population Registries Population-based disease registries, particularly of cancer and birth defects, are often considered the ideal source of cases. This is because the population at risk, whose identification is needed for control selection, is well defined by geographical or administrative boundaries. Practical limitations on their use include the speed with which cases can be identified and interviewed, to avoid selection bias from exclusion of those who may have died, and the feasibility of random sampling of controls.

Health Maintenance Organizations Large health maintenance organizations (HMOs) are advantageous as a source of cases for several reasons. The source population is enumerated and demographic data, as well as some exposure and covariate data, may already be available for everyone. This permits judicious selection of cases and controls using nested case-control, case-cohort or stratified two-phase sampling designs (Sect. 7.6). Relatively objective and inexpensive exposure assessments may be possible using routine medical or pharmacy records, some of which may already exist in electronic form. Similarly, cases are usually easily ascertained from reports of diagnoses within the organization. Of course, some assurance is needed that members of the HMO are unlikely to go elsewhere for diagnosis and treatment.

Hospitals and Clinics Historically, many case-control studies have been conducted using either a single or a small group of hospitals or clinics. This facilitates timely access to cases and increases the likelihood of their cooperation, thus limiting selection bias. On the other hand, definition of the source population from which the cases arose may be problematic, not to mention the practicality of obtaining random samples of controls from it.

7.4.1.3 Exclusion Criteria

In principle, any exclusion criteria may be used for cases so long as they are equally applied to the controls, and vice-versa, since they serve simply to restrict the source population. Thus, subjects may be excluded who reside in areas difficult to reach or who are not native speakers of the language of interview. Practical applications of this rule can be more subtle, however. Wacholder (1995) argues, for example, that exclusion of cancer cases who lacked a histologic diagnosis could inadvertently tend to exclude those from smaller, rural hospitals who were more likely to have exposures related to agriculture.

Exposure Opportunity Case-control studies are most informative when there is a substantial degree of exposure variability, so that the exposure is neither rare nor ubiquitous (Chase and Klauber 1965). Subjects known a priori to have no opportunity for exposure could be excluded on grounds of efficiency if the exposure was rare, since they would contribute little additional information. Thus, for example, women who were past reproductive age when oral contraceptives became popular should be excluded from a study of OC use and breast cancer (Wacholder et al. 1992a). On the other hand, since they provide valid information on the non-exposed, there is no logical basis for insisting that subjects without the opportunity for exposure should be routinely excluded from cohort and case-control studies (Schlesselman and Stadel 1987; Poole 1987).

7.4.2 Selection of Controls

7.4.2.1 Principles of Control Selection

Wacholder et al. (1992a) described three basic principles of control selection. The first two correspond roughly to considerations already developed regarding conceptual foundations and the use of matching. The third stems from the desire to minimize the effects of measurement error to which case-control studies are particularly susceptible.

The "Study-Base" Principle This is the principle that controls be randomly selected from disease-free members of the underlying cohort, also known as the source population (Kelsey et al. 1996) or study base (Miettinen 1985), at the times that cases are being ascertained (Sect. 7.2). When controls are in fact selected later, it sometimes mandates the random selection of a reference date for each control so that the distributions of the case diagnosis dates and control reference dates are comparable. Only exposures occurring prior to the reference/diagnosis date would be taken into account. This principle also implies that whatever exclusion criteria have been applied to the cases must also be applied equally to the controls.

The Deconfounding Principle This principle underlies the stratified sampling of controls to render possible, or improve the efficiency of, an adjusted analysis designed to control confounding (Sect. 7.3).

The Comparable Accuracy Principle This principle, controversial even in the authors' view, suggests that controls be selected so that the errors of measurement of their exposures and covariates are comparable to the measurement errors of the cases. The suggestion that dead controls be selected for dead cases, for example, is sometimes made on the basis of the comparable accuracy principle (Gordis 1982). Unfortunately, there is no guarantee that adherence to the principle will eliminate or even reduce bias (Greenland and Robins 1985). Unless the measurement error can be completely controlled, for example, by obtaining error-free measurements for a *validation* subsample of cases and controls and appropriately incorporating these data in the analysis, it can seriously compromise study validity even if case and control data are equally error prone (Sect. 7.5.3).

7.4.2.2 Sources of Controls

The appropriate source population for sampling of controls is determined by the study-base principal. When cases arise from an enumerated source population such as an HMO, controls may be sampled from this cohort using a nested case-control or case-cohort design (Figs. 7.1 and 7.2). One principal advantage of conducting epidemiological studies in the Nordic countries is their maintenance of national disease and population registers which may be exploited for case and control selection, respectively (see chapter ▸Use of Health Registers of this handbook). Standard survey sampling methods are often used to select controls for "population-based" studies in countries that do not maintain population registers. The most difficult and controversial problems of control selection arise with hospital-based studies.

Survey Sampling Methods for scientific sampling of populations have been developed by census bureaus and other government agencies throughout the world. The particular method most advantageous for any given epidemiological study will likely depend on the local administrative infrastructure. Survey sampling often proceeds in stages, where one first samples a large administrative unit, then a smaller one, and finally arrives at an individual household or subject. Such multi-stage "cluster" sampling introduces modest correlations in the responses of individuals sampled from the same primary sampling unit, more marked ones for individuals sampled from the same lower level cluster. Although often ignored by epidemiologists, usually at the cost of some underestimation of the variability in estimated relative risks, these correlations should be accounted for in a rigorous statistical analysis (Graubard et al. 1989). Fortunately, simple methods to accommodate cluster sampling are now routinely incorporated in the standard statistical packages.

Random Digit Dialing In view of the high costs of census bureau techniques in the United States, methods of survey sampling through the telephone exchanges have been developed (Waksberg 1978; Harlow and Davis 1988). Random digit dialing (RDD) has become increasingly popular for control selection in populations that have high rates of telephone access. Some implementations start with the telephone exchange of each case for sampling of controls that are thereby matched on somewhat ill-defined neighborhood factors (Robison and Daigle 1984). RDD methods may be costly for ascertainment of controls from minority populations, requiring dozens of calls to locate a suitable household (Wacholder et al. 1992b). They are particularly susceptible to bias because of higher selection probabilities for households that have more than one phone line or more than one eligible control and because of high rates of non-response (Sect. 7.5.1). The latter problem is likely to become increasingly serious in view of the persistent use of answering machines to screen out unwanted calls. The popularity of cell phones, moreover, eventually may make it infeasible to use RDD to draw a random control sample from a source population defined by geographical or administrative boundaries.

Neighborhood and Friend Controls Matched controls may also be selected from neighbors or friends of each case. For the former method, a census is taken of all households in the immediate geographical area of the case, and these are approached in a random order until a suitable control is found. Care must be taken to ensure that the control was resident at the same time the case was diagnosed. Even with these precautions, neighborhood sampling may yield biased controls for hospital-based studies since it will not be guaranteed that the control would have been ascertained as a case if ill, thus violating the study-base principle (Wacholder et al. 1992b). Neighborhood controls are also susceptible to overmatching due to their similarity to the cases on factors associated with exposure that are not risk factors for disease (Sect. 7.3.3). These same difficulties confront the use of friend controls, whereby a random selection is taken from among a census of friends provided by each case. There may be further selection on factors related to popularity since the friend selected as control may well not have listed the case as a friend had the friend become ill (Robins and Pike 1990). The primary advantage of friend controls would be a low level of non-response.

Hospital-Based Controls Many studies that ascertain cases through hospitals also select controls from these same hospitals, which is of obvious logistical convenience. Such controls are likely to have the same high response levels as the cases. The fact that they may be interviewed in a hospital setting, as the cases are, is an advantage from the perspective of the comparable accuracy principle (Mantel and Haenszel 1959). The major difficulties stem from the fact that the hypothetical study base, the *catchment* of persons who would report to the particular hospital if they developed the disease under study, may be different from the catchment population for other diseases. Furthermore, many of the disease categories from which controls could be selected may themselves be associated with the exposure. A large part of the planning of hospital-based case-control studies is devoted to the choice of disease categories thought to be independent of exposure and to have a similar catchment. The hope is that controls with such diseases will effectively constitute a random sample, vis-à-vis exposure, from the study base. Since the independence of exposure and disease diagnosis is rarely known with great certainty, a standard recommendation is to select controls having a variety of diagnoses so that the failure of any one of them to meet the criterion does not compromise the study (Wacholder et al. 1992b; Rothman and Greenland 1998, p. 101). If it is found later that a certain diagnosis is associated with exposure, those controls can be excluded.

7.4.3 How Many Controls per Case? How Many Control Groups?

Case-Control Ratios For a fixed number of study subjects, statistical power for testing the null hypothesis is optimized by having equal numbers of cases and controls. When the disease is extremely rare or acquisition of cases particularly expensive, however, it may be important and cost-effective to increase the numbers of controls. In order to have the same statistical power (to reject the null hypothesis of no exposure effect against local alternatives) as a design with equal numbers of cases and controls, a design with M controls per case would need only $(M + 1)/2M$ as many cases. When $M = 2$, for example, this would imply the use of three-fourth as many cases, but twice as many controls, to achieve the same power as a design with equal numbers. For a fixed number of cases, the relative efficiency of a design with M controls per case relative to one that uses an unlimited number of controls is therefore only $M/(M + 1)$. Since 80% of maximum efficiency can thus be obtained with $M = 4$, it is often inadvisable to seek a higher ratio. Exceptions occur when sampling and data collection for controls is substantially cheaper than for cases or if accurate estimation of large rate ratios, rather than a test of the null hypothesis, is the primary statistical objective (Breslow 1982; Breslow et al. 1983).

Multiple Control Groups Early case-control investigations, including the classic study of Doll and Hill (1952), often utilized two or more control groups. Indeed, multiple control groups were recommended by Dorn (1959) to improve the case-control study so that it would "provide a more valid basis for generalization." As explained by Hill (1971, pp 47–48), "If a whole series of control groups, for example, of patients with different diseases, gives much the same answer and only the one affected group differs, the evidence is clearly much stronger than if the affected group differs from merely one other group." Similar informal arguments have been put forward in favor of multiple control groups as a means of addressing the possible biases that may be associated with the use of any one of them (Ibrahim and Spitzer 1979). Working from a more formal perspective, Rosenbaum (1987) concluded that a second or third control group was useful *only* if supplemental information was available on whether such use addressed a specific bias. If controls sampled from separate sources have different exposure histories, even after statistical adjustment for potential confounders, this indeed suggests that similar adjustment of the case-control comparison may be inadequate to control confounding. However, failure to detect a difference among control groups may give a false sense of security unless they were deliberately selected to differ with respect to unmeasured potential confounders. Implementation of this last criterion would clearly require some guess as to what those unmeasured confounders might be.

Recent reviews of case-control methods have tended to shy away from the use of multiple control groups (Rothman and Greenland 1998, p. 106; Wacholder et al. 1992b). They argue that there is usually a single "best" control group and that since the discovery of an adjusted exposure difference with other control groups will force these to be discarded, the effort involved will have been wasted. However, there may not be a "best" control group, or its identification may be controversial. Discovery of a difference between control groups should generally encourage the investigator to seriously suspect that confounding may have compromised study results.

7.5 Pitfalls

Case-control studies are susceptible to the same biases and problems of interpretation that afflict all observational epidemiological studies. These include confounding, selection or sampling bias, measurement error, and missing data. Selection bias can be considered an extreme version of bias due to missing data where the entire observational record is missing for subjects who are in the source population but fail to be included in the study. Each of these topics is considered in detail in other chapters of this handbook. Many methods described there for dealing with such issues apply to case-control studies as well as to cohort studies. Attention is confined here to a few of the potential problems to which case-control studies are particularly susceptible.

7.5.1 Selection Bias

As elaborated at length in Sect. 7.2.1, the cases and controls in a case-control study are best viewed as resulting from outcome-dependent sampling from an underlying, often idealized cohort study. The goal is to estimate the degree of association of disease risk with exposure that would have been found had complete records been available for the entire cohort. The sampling of controls and sometimes even of cases may be stratified, for example, by sex and broad categories of age, but otherwise is supposed to be random within the subpopulations of diseased and non-diseased subjects. Selection bias arises when the sampling is in fact not random. It poses a major threat to the validity of case-control studies.

The effect of sampling bias is easy to demonstrate quantitatively for an exposure variable with two levels. For simplicity, we consider the effect on the odds ratio associating exposure with the cumulative risk of disease during a defined study period. The first 2 × 2 subtable displayed in Table 7.4 contains the population

Table 7.4 Effect of selection bias on odds ratio measures of association

	Population frequencies		Sampling fractions		Expected sample frequencies	
	Case	Control	Case	Control	Case	Control
Exposed	P_{11}	P_{10}	f_{11}	f_{10}	$f_{11} \times P_{11}$	$f_{10} \times P_{10}$
Non-exposed	P_{01}	P_{00}	f_{01}	f_{00}	$f_{01} \times P_{01}$	$f_{00} \times P_{00}$
Odds ratios	$\psi = \frac{P_{11} \times P_{00}}{P_{10} \times P_{01}}$		$\psi_{\mathrm{f}} = \frac{f_{11} \times f_{00}}{f_{10} \times f_{01}}$		$\psi^* = \psi_{\mathrm{f}} \times \psi$	

frequencies of subjects who are exposed and become diseased during the study period (P_{11}), who are not exposed and become diseased (P_{01}) and likewise the frequencies of being exposed or non-exposed and remaining disease-free (P_{10} and P_{00}, respectively). The target parameter of interest is the odds ratio ψ based on these population frequencies. As shown in the next two subtables, the odds ratio ψ^* expected from the case-control sample equals the product of the true odds ratio, ψ, times the cross products ratio of the sampling frequencies, denoted ψ_{f}. Hence $\psi = \psi^*$, that is, there is no bias, provided that $\psi_{\mathrm{f}} = 1$. This will occur when the sampling fractions for cases and controls are all the same, depend *only* on the disease outcome, that is, $f_{10} = f_{00}$ and $f_{11} = f_{01}$, or depend only on exposure, that is, $f_{01} = f_{00}$ and $f_{11} = f_{10}$. Often, the sampling fractions for cases are both near 1 whereas those for the controls are much smaller. The fact that this does not matter, provided that the sampling fractions for exposed cases and non-exposed cases are the same and similarly for controls, is another way of understanding why case-control studies provide estimates of the relative risk (disease odds ratio). Bias does occur when the sampling fractions depend *jointly* on exposure and disease, usually because exposed controls are more or less likely to be sampled than non-exposed controls. In a study that ascertained all the cases but sampled exposed persons as controls with twice the frequency as non-exposed persons, the estimated relative risk (odds ratio) would be twice the correct value. This is known as Berkson bias (Berkson 1946).

Some of the factors that contribute to selection bias are as follows:

Patient Dies Before Interview When cases are ascertained through a population based disease registry, a significant interval of time may elapse between initial diagnosis and notification to the registry. Some patients whose disease course is rapidly fatal may therefore not be interviewed in person, but are either excluded from the study or represented by a proxy interview subject to increased measurement error. This selection factor may affect both cases *and* controls in hospital based studies. It constitutes a major problem in reproductive epidemiology (see chapter ►Reproductive Epidemiology of this handbook).

Physician Refuses Consent Committees charged with protection of human research subjects may require that permission for participation be given by the patient's physician. This could affect control participation in hospital-based studies or case participation in general.

Subject Refuses Participation The most common reason for selection bias in case-control studies is refusal of the subject to participate, either actively by refusing to sign a consent form or passively by failure to return a questionnaire or turn up at the appointed hour for a laboratory examination. Cases with disease are often highly motivated to participate, whereas controls selected from the population are not. Unfortunately, control participation rates often depend on some correlate of exposure. Refusal rates for telephone surveys, for example, are higher for people who are older, have fewer social relationships, are less well educated, and have lower income (O'Neil 1979).

Subjects Ascertained Through Their Household Selection bias can occur when controls are ascertained by first contacting households to determine whether a control lives there who is suitable for matching to the case, and only a single control is selected from each household. In studies of childhood disease, where controls are matched on age within 2 years of the case, a child with a sibling in the same age range is less likely to be selected than one who has no such siblings (Greenberg 1990).

Random Digit Dialing Some other problems of selection bias are associated with the use of RDD for control ascertainment besides the fact that this method identifies households rather than individuals. Households without telephones stand no chance of selection, for example, whereas those with multiple telephones will be over-represented. The absence of a telephone may particularly affect minority populations.

7.5.2 Adjustments for Selection Bias in Study Design and Analysis

The most important consideration regarding selection bias is to avoid it so far as possible. At the design phase of the study, the exclusion criteria for both cases and controls may be chosen to maximize the probability of their ascertainment and participation. If RDD is used for control selection, this means taking the obvious step of excluding cases from households that lack telephones. Demographic, geographical, and linguistic factors may enter into the exclusion criteria for the same reason.

If selection bias cannot be avoided, as much data as possible should be gathered on *potential* case and control subjects to allow *prediction* of which of them go on to participate and which refuse. When sampling from the general population, it may be possible to use a recent survey of the same population for this purpose, provided of course that the survey itself had nearly complete response. If cases and controls are drawn from an enumerated population such as an HMO, data may already exist in medical or other records that can be used for this purpose.

At the time of analysis, one may attempt to adjust for selection bias in the same way that one adjusts for missing data. This is to use sampling weights for each participating subject, that is, those with "complete data," equal to the inverse predicted probability that the subject would have been selected given the data collected for this purpose at the design stage. This is only useful, of course, if there is substantial variability in the predicted probabilities. Alternatively, or additionally, one may statistically adjust the analysis for factors that are thought to be associated with selection but for which data are only available for participating subjects. Such adjustment would consist of stratification of the analysis on factor levels, or inclusion of the factor in a regression model for disease given exposure, just as one adjusts for confounders (Breslow and Day 1980, Sect. 3.8). However, if there is a substantial degree of non-response, it is quite unlikely that any adjustment will mitigate the serious biases that can result. There is simply no way to deal with it if selection fractions within factor levels used for adjustment purposes depend jointly on disease and exposure.

7.5.3 Measurement Error

A second major limitation of case-control studies is their susceptibility to measurement error. Cases and controls are often ascertained long after the relevant exposures have occurred. In spite of Dorn's (1959) admonition to use *objective* measures of exposure, most case-control studies of environmental risk factors continue today to measure exposure by interview or questionnaire. The potential for misclassification of exposure levels in such research is enormous. First, subjects may have only a vague memory of past exposures. Second, those who are diseased at the time of interview may recall these past events in a different way than those who are healthy controls. This may be in part because the early stages of their disease led to changes in behavior that made recollection of past practices more difficult. Interviewers may solicit and record answers differently if they have knowledge of the diagnosis or of the patient's status as case or control.

Austin et al. (1994) reviewed published reports of nine case-control studies of diet and cancer in which an attempt had been made to assess the accuracy of recall of dietary histories separately for cases and controls. According to their authors, three studies provided "weak" and four "moderate" evidence for recall bias. However, these results themselves were likely subject to measurement error and may have been understated in consequence.

Measurement error, whether or not it is differential between cases and controls, can compromise conclusions by seriously biasing the relative risk estimates from case-control studies that use dietary self-reports or similarly error-prone measurements. Prentice (1996) developed a mathematical model for measurement error that allowed for correlation of the error with the true exposure level and for systematic underreporting of exposure for persons with high exposure levels. He fitted the model to replicate measures of dietary fat intake, some taken using a 4-day food record and others using a food-frequency questionnaire for control subjects enrolled in the Women's Health Trial (Henderson et al. 1990). Employing results from international geographical correlation studies to generate the "true model," in which subjects at the 90th percentile of the distribution of dietary fat intake had three or four times the risk of disease as those at the tenth percentile, he showed that measurement error could plausibly reduce the relative risks to 1.1. The obvious conclusion from these calculations was that "dietary self-report instruments may be inadequate for analytical epidemiological studies of dietary fat and disease risk because of measurement error biases" (Prentice 1996).

A substantial and concerted effort has been made by statisticians to develop methods of data analysis that correct for the bias in relative risk estimates caused by measurement error (see chapter ▸Measurement Error of this handbook and the text by Carroll et al. 1995). Some require the availability of "gold standard," that is, error-free, measurements on a fairly large number of subjects in the validation subsample. Others assume that statistically independent true replicate measurements are available. Unfortunately, data collected in case-control studies rarely meet these stringent requirements, at least not in their entirety. It therefore behooves us to recall Bradford Hill's (1953, p. 995) sage advice:

> One must go and seek more facts, paying less attention to techniques of handling the data and far more to the development and perfection of the methods of obtaining them.

7.6 Conclusions

The case-control study played a major, successful role during the second half of the twentieth century in identifying risk factors for chronic disease. It has also proven helpful for evaluation of the efficacy of vaccination (Comstock 1994) and screening (Weiss 1994) programs. The twenty-first century will witness its continued use as a cost-effective study design, with increasing application in genetic epidemiology (Khoury and Beaty 1994) and particularly in the study of gene-environment interactions (Andrieu and Goldstein 1998). Statisticians and epidemiologists will continue to develop more efficient study designs and methods of data analysis that take full advantage of all available data. When a case-control study is conducted in an HMO, for example, some data will likely be available on either the exposure or the control variables for all subjects in the underlying cohort. *Two-phase* sampling designs, whereby *biased* samples of cases and controls are selected using the data available for all subjects, then offer the potential for much greater efficiency than the standard case-control design (White 1982; Breslow and Cain 1988; Langholz and Borgan 1995; Breslow and Chatterjee 1999). Chapter ▸Modern Epidemiological Study Designs of this handbook discusses these and other evolving study designs and analyses.

The advantages of case-control methodology in terms of speed and cost may have also contributed, ironically, to a diminished stature for epidemiology and biostatistics in the eyes both of the scientific community and of the general public (Breslow 2003). Part of the problem is an inherent aversion to the "black box" approach of risk factor epidemiology that associates cause and effect without the need for any understanding of pathogenetic mechanisms. Epidemiological findings are most convincing when supported by relevant laboratory research. Another part of the problem is the saturation of the news media with conflicting reports based on case-control and other studies that are too small, poorly designed, improperly analyzed, or overly interpreted. Taubes (1995) began his controversial and influential article on the limitations of epidemiology with the observation: "The news about health risks comes thick and fast these days, and it seems almost constitutionally contradictory." The epidemiologists he interviewed for this article cited the ability of confounding, selection bias and measurement error to overwhelm smaller exposure effects. One even suggested that no single study, no matter how well conducted, should be viewed as "persuasive" unless the lower limit of the 95% confidence interval for the rate ratio exceeded 3 or 4. Very few published studies, even when reported by the press as "suggestive" of an association, meet this stringent criterion.

Medical science and public health would be well served by fewer, larger case-control studies designed to test specific hypotheses that are carefully articulated in advance. Studies that can barely "detect" a relative risk of 2 may not provide convincing evidence of a dose-response gradient and are unlikely to enable one to determine whether an elevated relative risk in a particular disease subgroup, even one specified in advance, is evidence for the *specificity* of association that can be useful in causal interpretation (Weiss 2002). (There are of course exceptions, as when a unique exposure contributes to an outbreak of an extremely rare disease. Recall the DES-adenocarcinoma of the vagina story mentioned in the Introduction.) Investigators are also well advised to develop a strict protocol for selection of cases and controls and for collection and *analysis* of the data. Doll and Hill (1952) utilized such a protocol. They also had the advantage of working during the punch card era that discouraged "data dredging" and the inclusion of all but the most important variables in the analysis. A reasonable strategy might be to perform a maximum of three carefully planned analyses of the association between the primary exposure and disease: one without adjustment, one adjusted for a short list of confounders known a priori to be associated with disease, and the third adjusted for a specified list of known and suspected confounders. In case of conflict, the major interpretation would be based on the second analysis though the results of all three would be reported. Flexibility would be needed in application, of course, especially to accommodate changes in the study protocol after the study had commenced. Finally, investigators would be well advised to exercise greater caution in advertising their findings to the press before confirmation was forthcoming from other sources. By following basic principles of good statistical and scientific practice, the case-control study can gain credibility within the research community and enhance its standing as a basis for public health action.

Acknowledgements I am indebted to Sander Greenland, Noel Weiss, the editors and an anonymous referee for helpful comments on an earlier draft. This work was supported in part by grant R01-CA40644 from the US Public Health Service.

References

Aird L, Bentall HH, Roberts JAF (1953) A relationship between cancer of stomach and the ABO blood groups. Br Med J 1:799–801

Andrieu N, Goldstein AM (1998) Epidemiologic and genetic approaches in the study of gene-environment interaction: an overview of available methods. Epidemiol Rev 20:137–147

Armenian HK, Lilienfeld DE (1994) Overview and historical perspective. Epidemiol Rev 16:1–5

Armstrong RW, Armstrong MJ, Yu MC, Henderson BE (1983) Salted fish and inhalants as risk-factors for nasopharyngeal carcinoma in Malaysian Chinese. Cancer Res 43:2967–2970

Austin H, Hill HA, Flanders WD, Greenberg RS (1994) Limitations in the application of case-control methodology. Epidemiol Rev 16:65–76

Barlow WE (1994) Robust variance estimation for the case-cohort design. Biometrics 50: 1064–1072

Benichou J, Gail MH (1995) Methods of inference for estimates of absolute risk derived from population-based case-control studies. Biometrics 51:182–194

Berkson J (1946) Limitations of the application of fourfold table analysis to hospital data. Biom Bull 2:47–53

Breslow N (1982) Design and analysis of case-control studies. Annu Rev Public Health 3:29–54

Breslow NE (1996) Statistics in epidemiology: the case-control study. J Am Stat Assoc 91:14–28

Breslow NE (2003) Are statistical contributions to medicine undervalued? Biometrics 59:1–8

Breslow NE, Cain KC (1988) Logistic regression for two-stage case-control data. Biometrika 75:11–20

Breslow NE, Chatterjee N (1999) Design and analysis of two-phase studies with binary outcomes applied to Wilms tumor prognosis. Appl Stat 48:457–468

Breslow NE, Day NE (1980) Statistical methods in cancer research I: the analysis of case-control studies. International Agency for Research on Cancer, Lyon

Breslow NE, Lubin JH, Marek P, Langholz B (1983) Multiplicative models and cohort analysis. J Am Stat Assoc 78:1–12

Broders AC (1920) Squamous-cell epithelioma of the lip. A study of five hundred and thirty-seven cases. J Am Med Assoc 74:656–664

Carroll RJ, Ruppert D, Stefanski LA (1995) Measurement error in nonlinear models. Chapman and Hall, London

Chase G, Klauber MR (1965) A graph of sample sizes for retrospective studies. Am J Public Health 55:1993–1996

Cole P (1979) The evolving case-control study. J Chronic Dis 32:15–27

Comstock GW (1994) Evaluating vaccination effectiveness and vaccine efficacy by means of case-control studies. Epidemiol Rev 16:77–89

Cornfield J (1951) A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. J Natl Cancer Inst 11:1269–1275

Correa A, Stewart WF, Yeh HC, Santos-Burgoa C (1994) Exposure measurement in case-control studies: reported methods and recommendations. Epidemiol Rev 16:18–32

Cox DR (1972) Regression models and life-tables (with discussion). J R Stat Soc (Ser B) 34: 187–220

Daling JR, Weiss NS, Metch BJ, Chow WH, Soderstrom RM, Stadel BV (1985) Primary tubal infertility in relation to the use of an intrauterine-device. N Engl J Med 312:937–941

Doll R, Hill AB (1950) Smoking and carcinoma of the lung. Preliminary report. Br Med J 2:739–748

Doll R, Hill AB (1952) A study of the aetiology of carcinoma of the lung. Br Med J 2:1271–1286

Dorn HF (1959) Some problems arising in prospective and retrospective studies of the etiology of disease. N Engl J Med 261:571–579

Fleming PJ, Gilbert R, Azaz Y, Berry PJ, Rudd PT, Stewart A, Hall E (1990) Interaction between bedding and sleeping position in the sudden-infant-death-syndrome – a population based case-control study. Br Med J 301:85–89

Gordis L (1982) Should dead cases be matched to dead controls? Am J Epidemiol 115:1–5

Graubard BI, Fears TR, Gail MH (1989) Effects of cluster sampling on epidemiologic analysis in population-based case-control studies (Corr: V47 p. 779–780). Biometrics 45:1053–1071

Greenberg ER (1990) Random digit dialing for control selection – a review and a caution on its use in studies of childhood-cancer. Am J Epidemiol 131:1–5

Greenland S (1987) Estimation of exposure-specific rates from sparse case-control data. J Chronic Dis 40:1087–1094

Greenland S, Robins JM (1985) Confounding and misclassification. Am J Epidemiol 122:495–506

Greenland S, Thomas DC (1982) On the need for the rare disease assumption in case-control studies. Am J Epidemiol 116:547–553

Harlow BL, Davis S (1988) Two one-step methods for household screening and interviewing using random digit dialing. Am J Epidemiol 127:857–863

Henderson MM, Kushi LH, Thompson DJ, Gorbach SL, Clifford CK, Thompson RS (1990) Feasibility of a randomized trial of a low-fat diet for the prevention of breast-cancer – dietary compliance in the womens health trial vanguard study. Prev Med 19:115–133

Herbst AL, Ulfelder H, Poskanzer DC (1971) Adenocarcinoma of the vagina. N Engl J Med 284:878–881

Hill AB (1953) Observation and experiment. N Engl J Med 248:995–1001

Hill AB (1965) The environment and disease: association or causation? Proc R Stat Soc Med 58:295–300

Hill AB (1971) Principles of medical statistics. Oxford University Press, New York

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 47:663–685

Hsieh DA, Manski CF, McFadden D (1985) Estimation of response probabilities from augmented retrospective observations. J Am Stat Assoc 80:651–662

Ibrahim MA, Spitzer WO (1979) The case-control study: the problem and the prospect. J Chronic Dis 32:139–144

Jablon S, Neel JV, Gershowitz H, Atkinson GF (1967) The NAS-NRC twin panel: methods of construction of the panel, zygosity diagnosis, and proposed use. Am J Human Genet 19: 133–161

Kelsey JL, Whittemore AS, Evans AS, Thompson WD (1996) Methods in observational epidemiology, 2nd edn. Oxford University Press, New York

Khoury MJ, Beaty TH (1994) Applications of the case-control method in genetic epidemiology. Epidemiol Rev 16:134–150

Kupper LL, McMichael AJ, Spirtas R (1975) A hybrid epidemiologic study design useful in estimating relative risk. J Am Stat Assoc 70:524–528

Lane-Claypon JE (1926) A further report on cancer of the breast. Her Majesty's Stationery Office, London

Langholz B, Borgan O (1995) Counter-matching: a stratified nested case-control sampling method. Biometrika 82:69–79

Langholz B, Borgan O (1997) Estimation of absolute risk from nested case-control data. Biometrics 53:767–774

Langholz B, Goldstein L (1996) Risk set sampling in epidemiologic cohort studies. Stat Sci 11:35–53

Levin ML, Goldstein H, Gerhardt PR (1950) Cancer and tobacco smoking. A preliminary report. J Am Med Assoc 143:336–338

Liddell FDK, McDonald JC, Thomas DC (1977) Methods of cohort analysis: appraisal by application to asbestos mining. J R Stat Soc (Ser A) 140:469–491

Lilienfeld AM, Lilienfeld DE (1979) A century of case-control studies: progress? J Chronic Dis 32:5–13

Lin DY, Ying Z (1993) Cox regression with incomplete covariate measurements. J Am Stat Assoc 88:1341–1349

Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York

Lombard HL, Doering CR (1928) Cancer studies in Massachusetts. 2. Habits, characteristics and environment of individuals with and without cancer. N Engl J Med 198:481–487

MacMahon B, Cole P, Lin TM, Lowe CR, Mirra AP, Ravnihar B, Salber EJ, Valaoras VG, Yuasa S (1970) Age at first birth and breast cancer risk. Bull World Health Org 43:209–221

Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 22:719–748

Miettinen OS (1970) Matching and design efficiency in retrospective studies. Am J Epidemiol 91:111–118

Miettinen O (1976) Estimability and estimation in case-referent studies. Am J Epidemiol 103: 226–235

Miettinen O (1982) Design options in epidemiologic research: an update. Scand J Work Environ Health 8:7–14

Miettinen OS (1985) Theoretical epidemiology: principles of occurrence research in medicine. Wiley, New York

Neutra RR, Drolette ME (1978) Estimating exposure-specific disease rates from case-control studies using Bayes theorem. Am J Epidemiol 108:214–222

Neyman J (1955) Statistics – servant of all sciences. Science 122:401–406

O'Neil MJ (1979) Estimating the nonresponse bias due to refusals in telephone surveys. Public Opin Q 43:218–232

Poole C (1987) Critical appraisal of the exposure-potential restriction rule. Am J Epidemiol 125:179–183

Prentice RL (1986) A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 73:1–11

Prentice RL (1996) Measurement error and results from analytic epidemiology: dietary fat and breast cancer. J Natl Cancer Inst 88:1738–1747

Prince AM, Szmuness W, Michon J, Demaille J, Diebolt G, Linhard J, Quenum C, Sankale M (1975) A case-control study of the association between primary liver cancer and hepatitis B infection in Senegal. Int J Cancer 16:376–383

Robins J, Pike M (1990) The validity of case-control studies with nonrandom selection of controls. Epidemiology 1:273–284

Robins JM, Gail MH, Lubin JH (1986) More on 'biased selection of controls for case-control analyses of cohort studies'. Biometrics 42:293–29

Robison LL, Daigle A (1984) Control selection using random digit dialing for cases of childhood cancer. Am J Epidemiol 120:164–165

Rodrigues L, Kirkwood BR (1990) Case-control designs in the study of common diseases: updates on the demise of the rare disease assumption and the choice of sampling scheme for controls. Int J Epidemiol 19:205–213

Rosenbaum PR (1987) The role of a second control group in an observational study (with discussion). Stat Sci 2:292–316

Rothman KJ (1986) Modern epidemiology. Little, Brown, Boston

Rothman KJ, Greenland S (1998) Modern epidemiology, 2nd edn. Lippincott-Raven, Philadelphia

Schlesselman JJ (1982) Case-control studies. Oxford University Press, New York

Schlesselman JJ, Stadel BV (1987) Exposure opportunity in epidemiologic studies. Am J Epidemiol 125:174–178

Sheehe PR (1962) Dynamic risk analysis in retrospective matched pair studies of disease. Biometrics 18:323–341

Smith DC, Prentice R, Thompson DJ, Herrmann W (1975) Association of exogenous estrogen and endometrial carcinoma. N Engl J Med 293:1164–1167

Smith PG, Day NE (1984) The design of case-control studies: the influence of confounding and interaction effects. Int J Epidemiol 13:356–365

Smith PG, Rodrigues LC, Fine PEM (1984) Assessment of the protective efficacy of vaccines against common diseases using case-control and cohort studies. Int J Epidemiol 13:87–93

Taubes G (1995) Epidemiology faces its limits. Science 269:164–169

Thomas DC, Greenland S (1983) The relative efficiencies of matched and independent sample designs for case-control studies. J Chronic Dis 36:685–697

Tuyns AJ, Péquignot G, Jensen OM (1977) Le cancer de l'oesophage en Ille-et-Vilaine en fonction des niveaux de consommation d'alcool et de tabac. Bull Cancer 64:45–60

Wacholder S (1995) Design issues in case-control studies. Stat Methods Med Res 4:293–309

Wacholder S, McLaughlin JK, Silverman DT, Mandel JS (1992a) Selection of controls in case-control studies I. Principles. Am J Epidemiol 135:1019–1028

Wacholder S, Silverman DT, McLaughlin JK, Mandel JS (1992b) Selection of controls in case-control studies II. Types of controls. Am J Epidemiol 135:1029–1041

Waksberg J (1978) Sampling methods for random digit dialing. J Am Stat Assoc 73:40–46

Weinberg CR, Wilcox AJ (1998) Reproductive epidemiology. In: Rothman KJ, Greenland S (eds) Modern epidemiology, 2nd edn., Chap. 29. Lippincott-Raven, Philadeplphia, pp 585–608

Weiss NS (1994) Application of the case-control method in the evaluation of screening. Epidemiol Rev 16:102–108

Weiss NS (2002) Can the ‘specificity’ of an association be rehabilitated as a basis for supporting a causal hypothesis? Epidemiology 13:6–8

White JE (1982) A two stage design for the study of the relationship between a rare exposure and a rare disease. Am J Epidemiol 115:119–128

Wynder EL, Graham EA (1950) Tobacco smoking as a possible etiologic factor in bronchogenic carcinoma. A study of six hundred and eighty-four proved cases. J Am Med Assoc 143: 329–336

Yu MC, Ho JHC, Lai SH, Henderson BE (1986) Cantonese-style salted fish as a cause of nasopharyngeal carcinoma – report of a case-control study in Hong-Kong. Cancer Res 46: 956–961

Ziel HK, Finkle WD (1975) Increased risk of endometrial carcinoma among users of conjugated estrogens. N Engl J Med 293:1167–1170

Modern Epidemiological Study Designs 8

Philip H. Kass

Contents

P.H. Kass
Department of Population Health and Reproduction, University of California School of Veterinary Medicine, Davis, CA, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_8,

8.1 Introduction

A fundamental challenge pervasive to all experimental and non-experimental (observational) research is valid inference of causal effects. Although actions (through undefined mechanisms, but conventionally denoted by treatment, exposure, etc.) and reactions (e.g., disease, remission, cure) must occur by definition in individuals, the realm of epidemiology principally lies in the study of individuals in the aggregate, such as patients enrolled in clinical trials, participants at risk in cohorts, and populations. Until recently, innovations in epidemiological study design methods developed in the last half-century have hence largely fallen into the domain of the two major observational study designs used: cohort and case-control studies (cf. chapters. ►Cohort Studies and ►Case-Control Studies of this handbook).

The justification for these two designs has seemingly rested on their ability to approximate – albeit non-experimentally – the widely accepted paradigm of applied biomedical research: the randomized trial. But the potentially exquisite control that study investigators can often exercise to approach comparability of groups, and hence (in the absence of other biases) validity, cannot in general be commensurately achieved in non-experimental research. To compensate, some epidemiologists have invoked conventions such as the "study base" concept, which have intuitive appeal and some practical value in study design, but ultimately do little to contribute to an understanding of the theoretical underpinnings of observational studies. Whether one accepts randomized trials as models for epidemiological designs to emulate, or envisions study bases as natural referents, however, is immaterial, because these concepts neither add to the transparency of causal inference nor do they lend themselves to further advances in modern observational study design. Clearly and by necessity, a different paradigm is required.

The premise underlying such a paradigm is that individuals potentially live in a simultaneous duality of exposure and non-exposure, with a corresponding duality of observable and hypothetical outcomes. That individuals living under one exposure condition should, in theory, be compared with themselves under other counterfactual (i.e., counter to fact) conditions leads in turn to the premise of "case-only" studies. This chapter is thus predominantly concerned with those studies that juxtapose a case series (i.e., individuals that have experienced a singular health event) with a hypothetical comparison group. The legitimacy of such an approach, which implicitly precludes the need for studying an observable comparison (e.g., control) group, derives from the "potential outcomes" formulation of establishing causal inferences (Little and Rubin 2000). We proceed to show its utility in recent, and most probably future, developments in modern epidemiological study design.

Nevertheless, not all settings are conducive to comparing cases against hypothetical distributions. When these circumstances arise, such as when exposure effects are not intermittent and transient or when induction times are not transitory (enhancing the likelihood of carryover effects), traditional case-control studies and their more modern variants take on a greater relevance. This chapter will therefore also address two of these designs: case-cohort and cohort-centric case-control studies (better known as "nested case-control studies"). They share in common a conceptual understanding of case-control studies that is somewhat at odds with the more traditional view of these designs, namely, a de facto comparison of cases to non-cases. The latter viewpoint fails to account for the artificial nature of this dichotomy: non-cases can become cases, even during the course of a study. The issue is resolved by recognizing that incidence case-control studies are effectively cost- and size-efficient modifications of cohort studies, in which cases are not simply compared with non-cases, but instead with a sample from a cohort of individuals at risk, some of whom might become cases. So in reality, all case-control studies are cohort-centric, that is, "nested" within a cohort (though the constituents of cohorts are not always obvious and may defy enumeration, as in hospital-based case-control studies). Viewed in this unifying light, the case-control studies discussed in this chapter assume a rational connection (albeit with sometimes different validity assumptions) not only to cohort studies but in turn to case-only, experimental, and potential outcomes research as well.

8.2 Case-Cohort Studies

To illustrate the relationship between case-control studies with differing control selection strategies, consider a study period (t_0, t_1) within which cases occur and are recruited. The traditional view of case-control studies has been that of sampling controls from the population at risk at the termination of the study period (t_1). Although direct estimation of exposure-specific incidence proportions is not possible without ancillary information, employment of such cumulative incidence sampling of controls allows the use of the exposure odds ratio for estimation of the incidence proportion ratio (IPR) in a fixed population under the rare disease assumption (Cornfield 1951, chapter ▶Case-Control Studies of this handbook). An alternative control sampling approach frequently employed in nested case-control studies, *incidence density sampling*, allows odds ratio estimation of the constant incidence rate ratio without invocation of the rare disease assumption (Miettinen 1976). Interestingly, when the parameter of interest is the incidence proportion ratio, the matched odds ratio under density sampling even generally outperforms the exposure odds ratio under cumulative incidence sampling as a better estimator (Greenland and Thomas 1982).

Several authors (Kupper et al. 1975; Miettinen 1982; Prentice 1986) envision yet another type of case-control study in which controls are sampled exclusively from the cohort at risk at t_0, that is, prior to the onset of case occurrence in (t_0, t_1) and without regard to their future outcome status. This design, hereafter referred to as a case-cohort study (but also known as a case-base study for binary outcomes), is notable not so much for its dissimilarity with other case-control design variants, but rather for what it has in common. The key feature distinguishing an incidence density sampled (nested) case-control study from a case-cohort study is whether controls are matched to cases on time to outcome of the case (Wacholder and Boivin 1987). Thus, a case-cohort study can be likened to an *unmatched* nested case-control study, with controls randomly sampled from the population at risk at t_0 (hence excluding prevalent cases) without regard to failure times (a condition of being sampled is that all individuals are at risk at t_0). Still, a nested case-control study requires control sampling from the entire cohort at risk throughout (t_0, t_1), while the case-cohort study does not because controls are selected prior to any occurrence of incident cases. Exposure information, therefore, need only be obtained on those individuals sampled as controls at t_0 and any subsequent cases that may or may not be controls. This makes the study design particularly advantageous relative to a cohort study when collecting and processing of exposure information, such as laboratory samples, becomes expensive and time-consuming, and can hence be reserved for only the study subjects, with only a modest reduction in efficiency. Indeed, if the study is prospective, such processing can occur in advance of any cases. Another advantage of the case-cohort study is the ability to employ the same subset of the cohort as a control group for studies of multiple outcomes. In contrast, a nested case-control study requires different matched risk sets for each outcome studied.

When an outcome is rare (i.e., most observations are censored), follow-up of a full cohort, whether closed or dynamic, can be expensive and inefficient. In contrast, by sampling only a subset of the cohort, the case-cohort study affords advantages found in both cohort and case-control studies while offering considerable efficiency relative to both. Table 8.1 shows how a hypothetical case-cohort study from a cohort of N individuals is implemented when estimating the incidence proportion ratio is desired. In this example, an individual in the cohort has probability p of being randomly sampled (with sampling independent of exposure) when cohort membership is fixed at t_0. Thus, the *sub-cohort* is comprised of pN individuals on whom exposure information is ascertained (assuming no loss to follow-up). Note, however, that this principle can be extended to a dynamic cohort, allowing entrance at different times. Over the study period (t_0, t_1), a total of A exposed cases + B non-exposed cases occur, only some of whom $(pA + pB)$ are members of the sub-cohorts. The efficiency of the study is attributable to the economy of not evaluating

Table 8.1 Expected distribution of cases and controls in a case-cohort study with sampling fraction p, with sampling independent of exposure

	Exposed		Non-exposed		
	Sampled sub-cohort	Non-sampled remainder of cohort	Sampled sub-cohort	Non-sampled remainder of cohort	Total
Cases	$pA = A_1$	$(1-p)A = A_0$	$pB = B_1$	$(1-p)B = B_0$	M_+
Censored individuals	$p(N_1 - A) = C$	$(1-p)(N_1 - A)$	$p(N_0 - B) = D$	$(1-p)(N_0 - B)$	$N - M_+$
Individuals at risk at t_0	pN_1	$(1-p)N_1$	pN_0	$(1-p)N_0$	N

A number of exposed cases in cohort, B number of non-exposed cases in cohort, N_1 number of exposed individuals in cohort, N_0 number of non-exposed individuals in cohort, M_+ total number of cases in cohort censored individuals outside the sub-cohort (e.g., $(1-p)(N_1+N_0-A-B)$). It is noteworthy that cases from the entire cohort, not only those that are members of the sub-cohort, are utilized as cases in the study. Although it is advantageous to obtain a census of the cases, particularly when the outcome is rare, cases can potentially be sampled as well. A sub-cohort experiencing censoring over the study period can also be repopulated from the underlying cohort.

When the cohort is fixed, direct estimation of the crude *IPR* is possible without the rare disease assumption. Intuitively, with complete case ascertainment, one would expect the case-cohort odds ratio $(A/B)/(pN_1/pN_0)$ to estimate the *IPR*. However, Sato (1992a) develops a maximum likelihood estimator (MLE) that is asymptotically more efficient:

$$\widehat{IPR}_{\text{MLE}} = \frac{A\left[\frac{BM_1}{M_+} + D\right]}{B\left[\frac{AM_1}{M_+} + C\right]}, \tag{8.1}$$

where $A_1 = pA, A_0 = (1-p)A, B_1 = pB, B_0 = (1-p)B, C = p(N_1 - A)$, $D = p(N_0 - B)$, $M_1 = A_1 + B_1$, and $M_+ = A + B$ (see Table 8.1). It is important to note that this equation does not in general algebraically simplify further because, by design, control exposure information is obtained only on the members of the sub-cohorts N_1' and N_0', where $N_1' = pN_1$ and $N_0' = pN_0$, and not on the entire cohort N. The key difference between the two IPR estimators lies in how the number of sampled cases is employed. For the case-cohort odds ratio, the actual exposure-specific number of sampled cases (A_1 and B_1) is used in the calculation of the size of the cohort sample (i.e., $N_1' = A_1 + C$, and $N_0' = B_1 + D$). In contrast, the MLE estimates the exposure-specific number of cases as substitutes for A_1 and B_1 by multiplying the total number of cases in the exposed or non-exposed sub-cohorts (A or B) by the unconditional (on exposure) overall sampling fraction of cases (M_1/M_+, i.e., the proportion of all cases in the cohort originally chosen (i.e., prior to becoming cases) for inclusion in the combined exposed and non-exposed sampled sub-cohorts). Note that when the exposure-specific sampling fractions are equal to the overall sampling fraction, the two *IPR* estimators will be equal.

To illustrate these points, consider the data from a case-cohort study by Miettinen (1982) and cited by Sato (1992a). The study included ten individuals sampled in the exposed sub-cohort (N_1'), five of which became cases (A_1), and an additional five exposed cases occurred that were not in the sub-cohort (A_0). It also included 90 individuals sampled in the non-exposed cohort (N_0'), 15 of which became cases (B_1); an additional 35 non-exposed cases occurred (B_0). The intuitive estimator of the *IPR* for these data is 1.8, while the MLE is 2.2. In this example, the overall case sampling fraction, M_1/M_+, is $20/60 = 0.33$, and the exposed and non-exposed case sampling fractions are $5/10 = 0.5$ and $15/50 = 0.3$, respectively. However, if the exposure-specific sampling fractions of cases are identical, i.e., if the 35 non-exposed cases not in the sampled sub-cohort are changed to 15 non-exposed cases, then the exposure-specific case sampling fractions (0.5) are equal to the overall sampling fraction of cases, and both *IPR* estimates will equal 3.0.

When the outcome is not rare, random sampling of cases may be utilized. Sato (1992a) provides further details about incorporating such sampling into the analysis. When the outcome is rare, then few if any cases would be expected in the cohort sample ($M_1 << M_+$), and (8.1) reduces to the case-control odds ratio (AD/BC).

A large sample variance estimate of ln ($\widehat{IPR}_{\text{MLE}}$) (Sato 1992a) is given by

$$\widehat{\text{var}}[\ln(\widehat{IPR}_{\text{MLE}})] = \frac{1}{A} + \frac{1}{B} + \left[1 - 2\left(\frac{M_1}{M_+}\right)\left(\frac{1}{\frac{AM_1}{M_+} + C} + \frac{1}{\frac{BM_1}{M_+} + D}\right)\right] \qquad (8.2)$$

$$- \frac{N'^2 AB(A_0 + B_0)(A_1 + B_1)}{(A+B)^3\left(\frac{AM_1}{M_+} + C\right)^2\left(\frac{BM_1}{M_+} + D\right)^2},$$

where $N' = N_1' + N_0'$.

A $(1-\alpha)\%$ confidence interval for the crude *IPR* can be obtained from

$$\widehat{IPR}_{\text{MLE}} \exp\left\{\pm Z_1 - \alpha/2(\widehat{\text{var}}[\ln(\widehat{IPR}_{\text{MLE}})])^{\frac{1}{2}}\right\} \qquad (8.3)$$

where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution.

Calculation of the *IPR* estimator and its large sample distribution can be extended to stratified analyses using a Mantel–Haenszel estimator (Sato 1992b). With k strata, the estimator is given by

$$\widehat{IPR}_{\text{MH}} = \frac{\sum\limits_k \frac{N_{0k}' A_k}{T_k}}{\sum\limits_k \frac{N_{1k}' B_k}{T_k}}, \qquad (8.4)$$

where T_k is the total number of distinct individuals in stratum k ($A_k + B_k + C_k + D_k$).

The variance estimator of ln ($\widehat{IPR}_{\text{MLE}}$) that applies to both large strata and when data are sparse is

$$\widehat{\text{var}}[\ln(\widehat{IPR}_{\text{MH}})] = \frac{\sum\limits_k \frac{[(B_{0k}+D_k)A_k N_{1k}' + (A_{0k}+C_k)B_k N_{0k}' + A_{0k}D_k + B_{0k}C_k]}{(T_k)^2}}{\sum\limits_k \frac{N_{1k}' B_k}{T_k} \sum\limits_k \frac{N_{0k}' A_k}{T_k}}. \qquad (8.5)$$

As before, confidence limits can be obtained by applying the following formula:

$$\widehat{IPR}_{\text{MH}} \exp\left\{\pm z_{1-\alpha/2}\left(\widehat{\text{var}}\left[\ln(\widehat{IPR}_{\text{MH}})\right]\right)^{\frac{1}{2}}\right\}. \qquad (8.6)$$

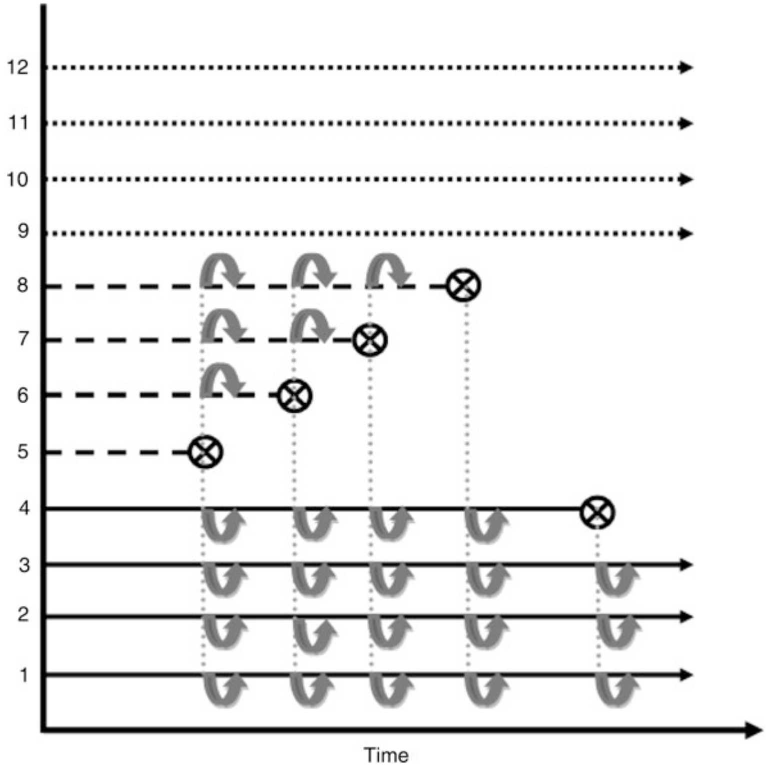


Fig. 8.1 Control sampling in case-cohort studies. Subjects 1–4 are selected for inclusion in the sub-cohort; subjects 5–8 are cases in the cohort (but not the sub-cohort); subjects 9–12 are members of the cohort who were not sampled for inclusion in the sub-cohort and did not become cases. *Circle* with X inside represents an incident case; *counterclockwise arrows* represent members of the sub-cohort used for the risk set at time of case occurrence; *clockwise arrows* represent non-members of the sub-cohort who were included in the study as cases and so were secondarily used as members of risk sets at times of earlier cases occurring

Additional details on crude and stratified incidence proportion ratio estimation can be found in Sato (1994).

When the parameter of interest is the incidence rate ratio, the analysis becomes more complex, and as noted earlier entrance into the cohort, and hence sub-cohort, can be dynamic. If all members of the cohort were followed and their exposure and outcome status (i.e., failure times) measured, a Cox proportional hazards regression model could easily be employed. Instead, a modification of this model is required for case-cohort data. Unlike a cohort study, at every time of case (regardless of sub-cohort membership) occurrence in the observation period, a risk set is created from the case and the uncensored members of the sub-cohort at that time (Fig. 8.1). For additional discussion about analytical issues in case-cohort studies, see Barlow et al. (1999) and Zhang et al. (2011).

Further developments in the methodology of dynamic case-cohort studies are addressing sampling approaches to make them even more efficient in weighing costs associated with exposure ascertainment and sampling cohort members and statistical precision. Breslow et al. (2009a) underscore the obvious point that conventional approaches to case-cohort studies fail to utilize available information that exists on cohort members not sampled. Sampling cases and non-cases within covariate strata, and then applying adjustments for stratum-specific weights, although not new (e.g., Walker 1982; White 1982) represents one area of development, as does incorporation of information from the non-sampled members of the cohort into analysis models. These approaches build upon the paradigm that case-cohort studies are de facto two-stage stratified sampling designs: the "first" stage recognizes the cohort as a representative sample of an unspecified target population, while the "second" stage involves sampling of cases and non-cases, either unconditionally or conditionally (stratified) on covariates, into the sub-cohort (Breslow et al. 2009a). Sampling weights for each individual are inversely related to their sampling probabilities in their strata, necessitating a weighted analysis that allows the stratum-specific contribution to equal what it would have been if all cohort members of the stratum had been analyzed. This weighting principle is illustrated using data from the Atherosclerosis Risk in Communities (ARIC) study (1989) reproduced in Breslow et al. (2009a), where sampling was stratified on age, sex, and ethnicity (Table 8.2). If additional confounding variables not utilized for stratified sampling are identified, the strata can be further subdivided using post-stratification.

Other methods of improving precision in the second sampling stage include calibration to adjust weights (Deville and Särndal 1992; Breslow et al. 2009a, b) and estimation to instead derive weights as inverse probabilities of sample selection from logistic regression (Robins et al. 1994). These improved weights enhance precision by utilizing covariate information available on the entire cohort. Additional information in implementing improving precision of regression coefficients can be found in Lin and Ying (1993) and Kulich and Lin (2004).

Although these methods have not yet been widely incorporated into commercial software, there are readily accessible options for implementing them. One based on the R language can be found at http://faculty.washington.edu/norm/IEA08.html

Table 8.2 Composition of cases and non-cases in the The Atherosclerosis Risk in Communities (1989) study population, stratified by ethnicity, sex, and age

	Non-coronary heart disease patients (non-cases)								Coronary heart disease patients (cases)
	Black				White				
	Female		Male		Female		Male		
	<55 Years	≥55 Years	<55 Years	≥55 Years	<55 Years	≥55 Years	<55 Years	?55 Years	
Stratum (k)	1	2	3	4	5	6	7	8	9
Cohort (N_k)	1,133	719	598	393	2,782	2,213	1,959	1,818	730
Sample (n_k)	59	54	42	71	88	154	117	147	604
Weights (N_k/n_k)	19.2	13.3	14.2	5.5	31.6	14.4	16.7	12.4	1.2

(Software IEA) (accessed on November 28, 2011), which utilizes the R-based "survey" package of Lumley (2008) at http://faculty.washington.edu/tlumley/survey/ (survey analysis in R) (accessed on November 28, 2011). A second based on R is the NestedCohort software of Mark and Katki (2006), available for download at http://dceg.cancer.gov/tools/analysis/nested-cohort (NCI Analysis Tools – NestedCohort) (accessed on November 28, 2011), with an accompanying tutorial at http://cran.r-project.org/web/packages/NestedCohort/vignettes/NestedCohort.pdf (accessed on November 28, 2011).

The choice between which type of case-control study to undertake – case-cohort or nested case-control – ultimately rests less on study-specific efficiency considerations than on what primary effect measure is required and whether the study is investigating one or multiple health outcomes. These issues are discussed in detail in Langholz and Thomas (1990) and Wacholder (1991).

8.3 Cohort Logic for Nested Case-Control Studies

Early incarnations of case-control studies owed their conceptual and inferential standing to Jerome Cornfield, who in 1951 demonstrates both that the exposure odds ratio could be estimated from a contrast of cases to controls (non-cases) and that under certain incidence assumptions (i.e., a low incidence proportion in a defined time interval in all groups of individuals with varying levels of exposure – and, by extension, within strata of controlled variables – hence the nominal "rare disease assumption"), it would closely approximate the incidence proportion ratio (or, as Cornfield calls it, the relative risk) in a fixed population. This conception of the case-control study, however, lacks a coherent linkage to the longitudinal structure of a cohort study and to the vision of a source population at risk from which cases arose. For a period of time in their formative years, case-control studies instead were sometimes fallaciously envisioned as directionally the opposite of cohort studies, leading to what later colloquially became the "trohoc fallacy."

The correspondence between longitudinal study designs became better understood and more widely accepted in the 1970s with the publication of the article *Estimability and estimation in case-referent studies* (Miettinen 1976). Here, Miettinen demonstrates that under certain designs involving control sampling from the population at risk, the exposure odds ratio is a consistent estimator of the incidence rate (density) ratio. Pointedly, unlike in the case-control design envisioned by Cornfield, no assumption of disease rarity is obligatory for estimation of the incidence rate ratio, a recognition that portended an even deeper understanding of control selection strategies to come.

Without yet defining specific examples, consider a population at risk followed over time; it is irrelevant whether this is retrospectively, ambispectively, or prospectively accomplished. The population may be closed or open; it may be completely enumerated or have its membership left uncharacterized; its distribution of exposure(s), baseline disease incidence, and relative effect of exposure (e.g., the incidence rate ratio) may be constant or time dependent. Regardless of which combination of these characteristics a population at risk possesses, the goal of the case-control study remains the same: an unbiased estimate of an epidemiological measure of impact (e.g., a synoptic incidence rate or risk ratio). This requires the exposure odds in the controls to equal (apart from random error incurred through sampling) the ratio of exposed relative to non-exposed time at risk in the population. Selection of controls to achieve this correspondence is accomplished through *incidence density* sampling (sometimes referred to as *density* or *risk set sampling*), where controls are sampled either directly from the population at risk or within strata of the population under observation, thereby providing a highly efficient way of studying a cohort without needing to obtain potentially expensive exposure information (e.g., abstracting medical records, genotyping, laboratory testing) on all members. The principle of sampling risk sets from this population provides a natural correspondence with failure-time cohort studies, which has spawned the name *nested* case-control studies. In practice, the word *nested* is used to distinguish case-control studies whose underlying cohorts are well defined from those whose members are not identified or ill-suited to enumeration. Controls, by virtue of their population membership, are at risk of becoming cases; conversely, individuals not at risk are ineligible to be controls because they are statutorily precluded from becoming cases in the study. Again, this underscores the subtle but important distinction between considering controls as individuals at risk as opposed to being strictly non-cases. Furthermore, controls remain members of the population at risk until the study period (i.e., the period of case acquisition) ends, and they are censored, until they exit the population or until they become diseased and become eligible to enter the study as cases. So long as they remain members, they are also eligible to be sampled in later risk sets as controls and are repeatedly counted. Just as cases in a cohort study no longer contribute time at risk to a cohort study, they are similarly no longer eligible to be selected as controls in a case-control study.

Figure 8.2, constructed primarily using Cornfield (1951), Greenland and Thomas (1982), and Greenland et al. (1986), summarizes the relationship between the exposure odds ratio from a case-control study under incidence density sampling and its corresponding measure of effect under the different scenarios discussed above (assuming that the effect of exposure can be synoptically represented in the absence of meaningful modification by other factors). Many scenarios depend upon the assumption that the distribution of exposure during the period of case accrual is invariant (exposure stationarity), although this assumption becomes less plausible if exposure is related to disease incidence and the disease is not rare in the source population, thus differentially depleting the exposed sub-cohorts. The latter authors also demonstrate that when the goal of a study investigator is to represent the effect of exposure through comparative risks, the exposure odds ratio derived under incidence density sampling of controls is a superior estimator of the incidence proportion ratio than the exposure odds ratio derived under cumulative incidence sampling (when controls are only taken at the end of the period of case acquisition and are hence ineligible to become cases).

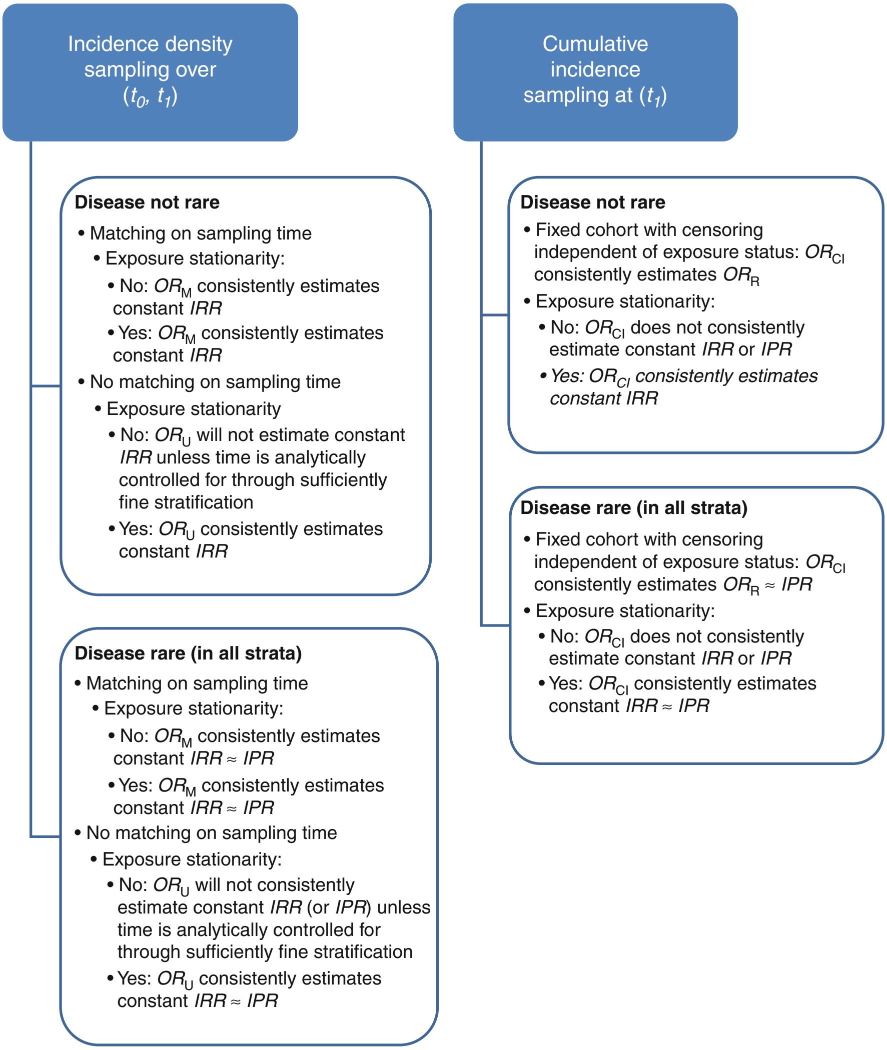

Incidence density sampling over (t_0, t_1)
Disease not rare
• Matching on sampling time
• Exposure stationarity:
• No: OR_M consistently estimates constant IRR
• Yes: OR_M consistently estimates constant IRR
• No matching on sampling time
• Exposure stationarity
• No: OR_U will not estimate constant IRR unless time is analytically controlled for through sufficiently fine stratification
• Yes: OR_U consistently estimates constant IRR
Disease rare (in all strata)
• Matching on sampling time
• Exposure stationarity:
• No: OR_M consistently estimates constant IRR ≈ IPR
• Yes: OR_M consistently estimates constant IRR ≈ IPR
• No matching on sampling time
• Exposure stationarity:
• No: OR_U will not consistently estimate constant IRR (or IPR) unless time is analytically controlled for through sufficiently fine stratification
• Yes: OR_U consistently estimates constant IRR ≈ IPR
Cumulative incidence sampling at (t_1)
Disease not rare
• Fixed cohort with censoring independent of exposure status: OR_{CI} consistently estimates OR_R
• Exposure stationarity:
• No: OR_{CI} does not consistently estimate constant IRR or IPR
• Yes: OR_{CI} consistently estimates constant IRR
Disease rare (in all strata)
• Fixed cohort with censoring independent of exposure status: OR_{CI} consistently estimates OR_R ≈ IPR
• Exposure stationarity:
• No: OR_{CI} does not consistently estimate constant IRR or IPR
• Yes: OR_{CI} consistently estimates constant IRR ≈ IPR

Fig. 8.2 Interpretation of odds ratios from case-control studies under two different control sampling schemes over the period of case accrual (t_0, t_1). *IPR* incidence proportion ratio, *IRR* incidence rate ratio, *OR* odds ratio, OR_{CU} odds ratio under cumulative incidence sampling, OR_M matched odds ratio under incidence density sampling, OR_R risk (incidence) odds ratio, OR_U unmatched odds ratio under incidence density sampling (Adapted from Cornfield (1951), Greenland and Thomas (1982) and Greenland et al. (1986))

Early recommendations for the ratio of controls to cases when exposure is constant relied on a simple formula under the null hypothesis (odds ratio = 1) for estimating the asymptotic relative efficiency of an unmatched case-control study relative to a cohort study, $r/(r+1)$, where r = the number of controls per case

(Breslow et al. 1983) in the crude situation where no confounders were analytically controlled. When $r = 0$, the study is trivially non-informative with respect to estimation of the odds ratio; when r becomes large, the case-control study efficiency approaches 1 and essentially approximates a cohort study. The relative improvement in efficiency and study power diminishes as r becomes larger, leading some to claim that there is little to be gained by having more than four controls per case in either a matched or unmatched design. Later study in control number sampling efficiency demonstrated the lack of universality of this formula (as demonstrated in Fig. 7.1 of Breslow and Day 1987) and the improved relative efficiency even with more than four controls per case for some combinations of odds ratio and prevalence of exposure in the source population. These authors showed that study efficiency conditional on the number of controls was also a function of the magnitude of the odds ratio and the prevalence of exposure in the source population, as well as whether or not matching was employed. In some instances, fewer than four controls per case would be satisfactory, while in others more than four would be indicated.

When exposure is not stationary (i.e., constant over time), as in administration of medicine, its effects can be both intermittent and transient. While it is typical to assess case and control exposure at the time of or immediately prior to the outcome (e.g., disease diagnosis or event), exposure in controls can more broadly be envisioned as a time-dependent variable divided into successive and distinct time windows. Such control individuals, representing the person-time at risk in the source population, would alternatively each have contributed exposed and non-exposed person-time in a cohort study.

Suissa et al. (2010) propose that sampling more than one of the control time windows (called control "person-moments") to improve study power when obtaining exposure information for multiple windows is more pragmatic than identifying more control individuals. The phrase "multitime case-control design" is thus coined to denote obtaining exposure information on all such time windows for which exposure information exists. Utilizing Suissa et al.'s notation, let k = the number of time windows, $k = 1, 2, \ldots, t$, with $k = 1$ representing the window closest in time to and t representing the window furthest in time from case occurrence. Let n = the number of cases and y_i, i = $1, 2, \ldots, n$ represent their binary exposure (0 = non-exposed, 1 = exposed), and m = the number of controls with x_{jk} representing their binary exposure (0 = non-exposed, 1 = exposed), with $j = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, t$. Thus, while a conventional case-control study would restrict use to n cases and m control time windows, there are no theoretical impediments to utilizing all n cases together with exposure measurements from mt control windows. The unadjusted synoptic exposure odds ratio under this design, $\widehat{OR}_{MT}$, is calculated as

$$\widehat{OR}_{MT} = \frac{a/c}{\sum_j B_j / \sum_j D_j}, \tag{8.7}$$

where $a = \sum_i y_i, c = \sum_i (1 - y_i), B_j = \sum_k x_{jk}$, and $D_j = \sum_k (1 - x_{jk})$.

Because exposures in successive time windows for each individual are unlikely to be independent, the variance of $\widehat{OR}_{MT}$ must be corrected as follows:

$$\widehat{\text{var}}\left(\ln(\widehat{OR}_{MT})\right) = \frac{1}{a} + \frac{1}{c} + V, \tag{8.8}$$

where $V = \Sigma((B_j/t) - R)^2/(R^2(1-R)^2 m(m-1))$ and $R = \Sigma B_j/mt$.

The above estimate assumes that the odds ratio is invariant across duration of exposure, which allows exchangeability (and potentially random sampling) of control exposure time windows (Suissa et al. 2010). If exposure did not vary over time, then the dependency between control person-moments would lead to complete concordance (i.e., the intraclass correlation = 1), and there would be no efficiency gain by sampling more person-moments; instead, more controls would be needed. If, in contrast, exposure was free to vary between person-moments, as in time-dependent exposures of transient duration, then it may be advantageous to sample multiple within-subject control intervals instead of acquiring more controls. For example, Suissa et al. note that with complete independence between person-moments, the improvement in precision by having five versus two person-moments is roughly equivalent to having four versus one control individuals. Furthermore, this approach is suited to acute exposures of limited duration, but not for those whose effects are cumulative (where each control subject would only have a single exposure time window). It also holds promise for extension to regression models that correct for the within-control correlation over time and can adjust for other covariates.

As with case-cohort studies, two-stage sampling for incidence density sampled case-control studies has been proposed to improve study efficiency (Breslow and Chatterjee 1999). To demonstrate its justification, these authors utilize data from the National Wilms' Tumor Study Group (D'Angio et al. 1989) shown in Table 8.3. The table contains 4,088 embryonal kidney cancer patients who either did (cases) or did not (controls) relapse, cross-classified by initial (institutional) tumor histological finding of prognostically favorable or unfavorable. The authors sought to validate the initial histology due to concerns about misclassification but wanted to do so without reassessing every tumor. A probability sample of the entire patient cohort would yield an unacceptably low number of cases while capturing an unnecessarily high number of controls. An alternative approach would be to do

Table 8.3 Data from National Wilms' Tumor Study Group, reproduced from Breslow and Chatterjee (1999)

	Entire cohort		Case-control study		Weighted case-control study	
Histology	Cases	Controls	Cases	Controls	Cases	Controls
Favorable	415	3,262	415	536	415	316
Unfavorable	156	255	156	35	156	255
Total	571	3,517	571	571	571	571
Odds ratio	4.8		5.8		0.47	

a nested case-control study, with all cases and a sample of controls to achieve a 1:1 ratio. This approach is scarcely satisfactory, however, due to the relatively small number of controls ($n = 35$) whose tumor histology was unfavorable. By performing weighted sampling that constrained the number of controls to remain the same ($n = 571$) while including all controls with unfavorable histology, the authors were better able to achieve a "balance" between the favorable ($n = 316$) and unfavorable ($n = 255$) sub-cohorts.

The odds ratio in the weighted case-control study is clearly biased, reflecting that the weighted sampling violated one of the central tenets of control selection: that the probability of control selection through sampling is statistically independent of exposure distribution. Of course, matching is an obvious counterexample, but the selection bias is easily remediated through appropriate analytical methods. As with case-cohort studies, weighting reflects the probability of being sampled in stage two conditional on stratum defined by one or more covariates and that an individual was sampled in stage one. Upon analysis, Breslow and Chatterjee (1999) find very close concordance between the regression coefficients and standard errors from the entire cohort and the weighted case-control study while only needing to validate histology on 571 of 3,517 (16%) of controls.

Special survey software must be applied to remediate the bias. Many software packages (open source and commercial) provide options for survey analysis. These include Lumley's "survey" package at http://faculty.washington.edu/tlumley/survey/ (survey analysis in R) (accessed on November 28, 2011). A recent review by Witt (2009) examines software commercially available to provide sample weight adjustments.

Another approach to two-stage sampling of controls in risk sets for improving efficiency is proposed by Langholz and Borgan (1995) who recognize that when exposure is rare in the cohort, randomly sampling controls would be unlikely to yield precise measures and that matching cases and controls on one or more confounders can lead to matched sets concordant on exposure that do not contribute to exposure effect estimation and preclude confounder effect estimation (although not exposure effect measure modification by the stratification variable). They develop a control sampling method called *counter-matching* to improve discordancy by maximizing exposure variability within risk sets. While the conventional notion of individual matching in a case-control study is to make controls as alike as possible to the cases on the matching factors, the opposite holds in counter-matching. Although the sampling mechanism is counterintuitive because control selection is not independent of exposure, case-control studies can be made substantially more efficient under certain conditions with proper analysis that controls for the deliberately weighted sampling (Langholz and Clayton 1994).

Consider a nested case-control study in which the goal is to estimate the effect of an exposure with two or more levels on the incidence of a binary outcome (e.g., disease status). If exposure is difficult to obtain or expensive to measure (stage two information), it would be wasteful of human and/or financial resources, if not outright infeasible, to obtain exposure information from an entire cohort when only a fraction of it will ultimately be selected as controls. However, there may exist a readily available surrogate for exposure (stage one information) from the entire cohort whose measurement may provide a useful although imperfect substitute, for example, utilizing readily obtainable duration of mining employment as a surrogate for cumulative silica dust exposure (Steenland and Deddens 1997) or assessing whether an individual received any vaccines in a defined time period as a substitute for knowing if a specific vaccine brand or lot (the exposure) suspected of causing an adverse event was administered. Alternatively, there may be complete ascertainment of exposure on all members of the cohort, but detailed information on confounders essential for study validity may be only cost-effective to obtain once controls have been selected for risk sets. For example, radon exposure histories from the Beaverlodge Uranium Miners Cohort of over 10,000 miners were obtained to assess its association with lung cancer mortality, while detailed smoking histories were selectively obtained from next of kin on only 46 cases and 95 controls who died between 1950 and 1980 (L'Abbé et al. 1991).

The procedure for performing counter-matching can be understood by first considering a nested pair-matched (1:1 design) case-control study and a dichotomous exposure. Each case in a surrogate exposure stratum is matched to a single control in a surrogate non-exposure stratum (i.e., the *opposite* stratum of the case), and a case in a surrogate non-exposure stratum is matched to a single control in a surrogate exposure stratum. By extension, more than one control per risk set can be accommodated for multiple strata (Langholz and Clayton 1994; Langholz 2005). In the matched case of one case per risk set (individual matching), suppose the counter-matching factor, known on all risk set members, has r strata, $r = 1, 2, \ldots, R$, with the case falling into only one stratum. The cohort is comprised of N individuals divided into the R strata, with n_r individuals in each stratum from which m_r individuals are sampled without replacement. The variability in exposure (or its surrogate) is realized by taking m_r controls from each stratum *except* the one containing the case; for that stratum $m_r - 1$ controls are sampled (Table 8.4). In an unmatched study (i.e., multiple cases (d_r) occupying R strata), the number of controls per counter-matched stratum will be determined by the number of cases, that is, $m_r - d_r$.

Table 8.4 Illustration of counter-matching using one case matched to m_r controls (Langholz 2005)

	Sampling stratum					
	1	2	3	…	R	Total
Cases	0	0	1		0	1
Controls	m_1	m_2	$m_3 - 1$		m_R	$\Sigma_r m_r - 1$
Total sampled in risk set	m_1	m_2	m_3		m_R	$\Sigma_r m_r$
Total eligible in cohort for risk set	n_1	n_2	n_3		n_R	$\Sigma_r n_r$
Weight	n_1/m_1	n_2/m_2	n_3/m_3		n_R/m_R	

To obtain unbiased estimators, counter-matched nested case-control studies must correct for selection bias induced by the sampling design. This is accomplished by utilizing the log weights (i.e., ln (n_j/m_j)) as an offset in a conditional logistic regression analysis. Unlike conventional case-control matching, where main effects of matching factors are not estimable, counter-matching allows direct estimation of the effects of the stratification factor, as well as its interaction with other model variables.

When counter-matching on an exposure or its surrogate, a nested case-control study can be more efficient than one in which matched controls are randomly sampled from a risk set. However, it is important that the counter-matching variable is correlated with the variables of interest for the analysis, be they precise exposure measures or confounders. Thus, counter-matching on a rare exposure can make a study much more efficient if the goal is to estimate an exposure effect. To see this, in the special case of a dichotomous (+ = present, − = absent) exposure (E) and counter-matching covariate (C), consider the sensitivity (Se) and specificity (Sp) of the covariate (P($C+|E+$) and P($C-|E-$), respectively) for predicting true exposure. When exposure and disease are independent, the asymptotic relative efficiency of a 1:1 counter-matched versus a 1:1 randomly sampled nested case-control study is given by (Langholz 2005):

$$2[SeSp + (1 - Se)(1 - Sp)]. \tag{8.9}$$

Thus, the efficiency gained from counter-matching arises when both Se and Sp are either greater or less than 0.5. Additional asymptotic relative efficiencies for 1:1 and 1:3 matching under exposure prevalences of 5% and 50% when the sample-based incidence rate ratio = 2 are provided in Table 1 of Langholz and Borgan (1995) when counter-matching on a surrogate of exposure. In general, relative efficiency rises as sensitivity and specificity both rise, but specificity is more influential than sensitivity, particularly as the effect of exposure increases; relative efficiency is also higher for rare rather than common exposures. When counter-matching on true exposure and controlling for a confounder, counter-matching results in substantial gains in efficiency for various combinations of exposure effects and confounder-exposure associations, particularly when exposure is rare and the effect of exposure increases (Table 2 of Langholz and Borgan 1995; Table 3 of Langholz and Clayton 1994). As a practical matter, Steenland and Deddens (1997) find in their silicosis study that counter-matching with three controls had an equivalent efficiency of randomly sampling ten controls.

In contrast, the advantages of counter-matching on exposure (or its surrogate) fail to extend to counter-matching on a confounder. Langholz and Clayton (1994) demonstrate in their Table 4 such deleterious consequences: when both the confounder and exposure are rare, the efficiency of the counter-matched study is often profoundly less efficient than a nested case-control study unless the association between the confounder and exposure is very strong (i.e., when the confounder becomes a de facto surrogate of or proxy for the exposure).

8.4 Case-Crossover Studies

8.4.1 Introduction

Under circumstances that could be described as both ideal and impossible, it would become a trivial matter to evaluate causation within an individual, thus obviating the need for observational studies at all. This ideal, which is assumed in Table 8.5, reflects the counterfactual capacity to evaluate not only the health or disease experience of a cohort of exposed individuals (column 1) but what the experience of this cohort would have been had it been possible to evaluate these same individuals during the same time frame as when they were actually exposed, with exposure hypothetically being removed or its effect completely blocked (column 2).

Such a comparison between these same individuals under two identical conditions (save for the fact of exposure) could then lead not only to an individual-by-individual assessment of the causal or preventive effect of exposure but also the average causal effect of exposure in the exposed *cohort (A/A*)*. The value of the non-exposed cohort (column 3) for comparison with the exposed cohort (column 1), and hence effect estimation, lies in the validity of its exchangeability property, that is, $B/N_0 = A^*/N_1$.

Given the pragmatic limitations which epidemiologists operate under – pointedly, the inability to have study participants and investigators reexperience events in time, rendering column 2 unobservable – it becomes a challenge, if not an imperative, to design studies to come as close as possible to the ideal. In experimental research, the crossover study, in which individuals crossover between periods of different exposures (possibly including non-exposure, as in a control, placebo, or sham treatment), affords the opportunity to observe different treatments within the same individual, albeit at different (and potentially confounding) time periods. The allure of this experimental approach, underlying the ability to control for individual-level confounders, is offset by the assumption that neither period nor carryover effects occur. Thus, this design is optimally suited to exposures with both a rapid onset of action as well as a brief effect period.

Table 8.5 Hypothetical incidence data from exposed (as well as counterfactually non-exposed) and non-exposed cohorts

		Exposed*	
Diseased	A	A^*	B
Non-diseased	C Exposed	C^*	D Non-exposed
Total person-time at risk	N_1	N_1	N_0
Incidence proportion	A/N_1	A^*/N_1	B/N_0

Note that * indicates the same exposed group when counterfactually non-exposed. Therefore, A+C is constrained to equal A*+C*

The case-crossover design (Maclure 1991) is the observational study analog of the crossover study. A key distinguishing characteristic of the former is that study participants – not investigators – non-experimentally determine their own exposure. A very important feature of case-crossover studies, like crossover studies, is that they are apposite to episodic exposures with short induction and transient effect periods for uncommon events. Thus, it is possible to envision, following an exposure that modifies risk in individuals, a brief induction period during which risk does not change, followed by a period subdivided into intervals characterized by varying degrees of altered risk, and ultimately a return to baseline risk. The time interval of elevated or decreased risk following exposure is known as the effect period. In practice, the effect period may not be known, but may be inferred from ancillary information (e.g., pharmacokinetic properties of a drug or environmental agent, duration of effect of endogenous catecholamine release following physical exertion, period of heightened immune activity following vaccination, half-life of a chemical, incubation times of a pathogen following infection, etc.). Sensitivity analyses postulating different combinations of induction time and duration of exposure effects can provide insights into periods of maximum influence. As noted by Maclure (1991), postulating effect periods that are either too brief or too long leads to non-differential misclassification of exposure, resulting in effect measures that are biased toward the null. In the absence of other biases, the optimal choice of an effect period is that which minimizes non-differential misclassification and hence maximizes the effect measure.

8.4.2 Case-Control Logic for Case-Crossover Studies

The simplest conception of a case-crossover study envisions, for every subject, a contrast of the duality of states: a comparison of their exposure immediately prior to their event time (the event period) to their exposure at an *antecedent* (i.e., non-event) time. This bears closely analogy to a pair-matched case-control study in which a case's exposure is compared to a control's exposure, but with subtle differences. All individuals in a case-crossover study are, by definition, "cases" because they experienced the event, so the exposure at the time of the event substitutes for the case's exposure in a case-control study, and the exposure at the antecedent time – when the individual is still at risk of the event – substitutes for the control or referent in a case-control study. To avoid confusion and reinforce the case-only design model, non-event comparison times are heretofore referred to as *referent* times. Taking referent times exclusively from the time window prior to the event has the advantage of ensuring that the event itself does not affect the exposure measurements (i.e., as in "reverse causation"), but there are also disadvantages that will be discussed below.

Although the case-crossover construct is conducive to controlling within-individual time-independent confounding, the choice of referent times preceding the event period must judiciously be made when time-dependent variation in exposure is plausible (i.e., when the assumption of a constant distribution or stationarity of exposure across referent times is unlikely to be met). Just as matched case-control studies can achieve greater efficiency when the number of controls per case increases, potentially increasing the number of discordant case-control sets for analysis, more than one referent time can be selected for every event time. The injudicious use of referent times in the absence of exposure stationary can lead to a case-crossover form of selection bias analogous to that of a case-control study (Greenland 1996). This can especially be manifest when referent times are taken long before the event time, even in the absence of time (e.g., seasonal or weekly) trends.

Mittleman et al. (1995) examine different approaches to modeling case-crossover data from the control period, and the relative sampling efficiencies and restrictiveness of the assumptions inherent in each approach. Four of the modeling approaches involved different methods of studying the control period in the 25 one-hour periods immediately preceding myocardial infarction, with exposure being heavy exertion. The first approach the authors evaluate is called the "Pair-Matched Interval Approach," in which the one-hour hazard period immediately prior to the onset of myocardial infarction is contrasted with the one-hour control period from the same time of day 24 h earlier within the same subject. These data control for confounding by time of day, without the need to specify the baseline hazards in each of the day's 24 one-hour periods. The analysis can be made more statistically efficient by increasing the number of control intervals sampled. However, this method is not well suited to evaluating other exposures that occur at regular and predictable intervals during a 24-h period. For example, if an individual reliably self-administers a medication with potential cardiovascular effects to be taken at 08:00 and 20:00 h each day, then by comparing exposure on the day of disease occurrence to the same time 24 h earlier, this approach will ensure perfect concordance between hazard and control periods, resulting in bias toward the null. A second strategy, the "Nonparametric Multiple Intervals Approach," involves explicitly modeling exposure information on each of the 24 one-hour categorical intervals in the day prior to the myocardial infarction. This model, besides controlling for confounding by time of day, assumes a homogeneous exposure effect on each of the 24 hour-specific times (i.e., time is not an effect measure modifier). Although this categorical model, with indicator variables representing the one-hour intervals, makes no assumption about the functional relationship between the time of day and myocardial infarction incidence, a third model, the "Parametric Multiple Intervals Approach" does exactly that. Instead of indicator variables, the authors employ sine and cosine functions, based on consistency with prior (external) information on temporal patterns of myocardial infarction occurrence. Under the assumption that the model is correctly specified, it provides a synoptic estimate of the effect of exposure while controlling for confounding by time of day.

Maclure (1991) illustrates another referent selection strategy using case-control logic by assuming that some exposures (e.g., behavioral) occur in individuals with predictable frequency (note that many such exposures would be highly susceptible to modification following the event). In this "Usual Frequency Approach" a time window prior to the event is subdivided into time within or outside exposure effect

Table 8.6 Representation of case-crossover data for a single individual developing an outcome event either within or outside an exposure effect period

	Time		
	Within effect period	Outside effect period	Total
Case occurrence	a_i	b_i	
Person-time	N_{1i}	N_{0i}	T_i

periods. Assuming the absence of a time trend in exposure within the window, the referent person-time exposure odds can be contrasted with the exposure odds at the event time for effect estimation. Data from an individual i in a "Usual Frequency Approach" case-crossover study can be envisioned in Table 8.6. When the distribution of exposure is stationary over the case and control period and the outcome event is uncommon, the Mantel–Haenszel incidence rate ratio ($\widehat{IRR}_{\text{MH}}$) is approximately asymptotically unbiased (Vines and Farrington 2001). Each case occupies a unique stratum. $\widehat{IRR}_{\text{MH}}$, corresponding to the average proportionate change in the rate of the outcome resulting from exposure, can be calculated from the following formula (note: $N_{1i} + N_{0i} = T_i$):

$$\widehat{IRR}_{\text{MH}} = \frac{\sum_i a_i N_{0i} / T_i}{\sum_i b_i N_{1i} / Ti} \tag{8.10}$$

The variance of $\ln(\widehat{IRR}_{\text{MH}})$ (Greenland and Robins 1985) can be estimated by

$$\widehat{\text{var}}[\ln(\widehat{IRR}_{\text{MH}})] = \frac{\sum_i (a_i + b_i) N_{1i} N_{0i} / T_i^2}{\left(\sum_i a_i N_{0i} / T_i\right) \left(\sum_i b_i N_{1i} / T_i\right)}. \tag{8.11}$$

To illustrate, consider the hypothetical data in Table 8.7 (adapted from Table 4 of Maclure 1991). The column of referent exposure odds is equivalent to Maclure's expected concurrence odds and refers to the ratio of the expected amount of time an individual spends in the effect period, based on the usual frequency of exposure, to the expected amount of time an individual spends outside the effect period (N_{1i}:N_{0i} from Table 8.6) over the duration of retrospective follow-up for cerebrovascular accidents. In this example, the period of retrospective follow-up was 6 months, or equivalently 4,383 h. For example, for participant 1, the usual frequency of aerobic exercise was three times per week. If the effect period is assumed to be 2 h beginning immediately after the cessation of exercise, then the expected number of hours spent in the 6-month effect (exposure) period = 2 h/day × 3 days/week × 26 weeks = 156 h (N_{1i} in Table 8.6), leaving $4,383 - 156 = 4,227$ h non-exposed (N_{0i} in Table 8.6). Because participant 1s cerebrovascular accident occurred within

Table 8.7 Data from a hypothetical case-crossover study evaluating the relationship between aerobic exercise and the onset of cerebrovascular accident. The effect period is 2 h. The period of retrospective follow-up is 6 months or equivalently 4,383 h

			Exposure odds	
Individual (i)	Usual frequency of exposure	Last exposure before cerebrovascular accident	Event time (a_i:b_i)	Referent (N_{1i}:N_{0i})
1	3/week	30 min	1:0	156:4,227
2	1/week	1 day	0:1	52:4,331
3	1/month	21 days	0:1	12:4,371
4	0/month	1 year	0:1	0:4,383
5	5/week	45 min	1:0	260:4,123
6	0/month	2 years	0:1	0:4,383
7	2/month	15 h	0:1	24:4,359
8	4/week	3 h	0:1	208:4,175
9	1/week	6 h	0:1	52:4,331
10	0/month	4 years	0:1	0:4,383

the 2-h effect period, the observed exposure odds for this individual – the ratio of the a$_i$ and b$_i$ cells in Table 8.6 is 1:0. Using (8.10) and (8.11), the Mantel–Haenszel incidence rate ratio for cerebrovascular accidents within 2 h of aerobic activity in this sample of individuals and its corresponding 95% confidence interval are 24.0 and 3.1–188.1, respectively.

The preceding example derived the usual frequency of exposure by utilizing a census of the information from the 6 months prior to the event onset. In practice, retrospective follow-up time may be briefer or longer, depending on the stability of the exposure distribution. The analysis presupposed that no within-individual confounding occurred over time and that the effect of aerobic exercise did not depend on time of day.

In all the above models, it was possible to estimate a single summary effect of exposure under the assumption of no effect modification of the exposure-outcome association by time of day. Mittleman et al. (1995) show that while the number of control periods sampled has little effect on the incidence rate ratio estimate, as the periods sampled increase, the relative efficiency (marked by a narrowing of the respective confidence intervals) concomitantly increases as well, regardless of what underlying model assumptions are invoked. The "Usual Frequency Approach" leads to the smallest variance of the estimated logarithm of the incidence rate ratio, and hence narrowest confidence interval, but at the cost of assuming no within-individual confounding unless further information is available on the complex conditional relationships among all determinants of risk. Although the modeling approach by these authors may not be appropriate in all circumstances, they do underscore the point that case-crossover studies, like their case-control counterparts, can be conducted using different control period sampling schemes and employing different exposure and confounder/effect modifier assumptions.

8.4.3 The Exchangeability Assumption

Already noted is the potential for bias arising from time trends in exposure when referent periods are constrained to occur prior to the period (Greenland 1996). A more subtle source of bias arises when exposures measured from different referent periods within individuals are not independent (Vines and Farrington 2001). Recall that in a case-control study's matched set comprised of one case and two or more controls, there is no natural ordering of controls, so their respective exposures are *exchangeable* in the sense that switching exposures among the controls would have no impact on odds ratio estimation. In contrast, multiple referent periods and the event period are constrained to be ordered in case-crossover studies because they are deliberately selected to occur at fixed times prior to the event (i.e., using the case-control logic cited above). The assumption of exchangeability therefore need no longer be valid because of the potential for exposures within individuals to be correlated, and hence non-independent, between different time intervals. A necessary condition for the valid use of the Mantel–Haenszel estimator is that pairwise exchangeability exists between the event period and each referent period; that is, the probability of exposure at the event time and non-exposure at each referent time is equal to the probability of non-exposure at the event time and exposure at each referent time. Pairwise exchangeability is a less restrictive assumption than global exchangeability, which requires exposure independence between all time periods within matched sets in a case-crossover study. Such lack of global exchangeability in a case-crossover study can lead to biased odds ratio estimates when utilizing conventional multivariate approaches for matched data, such as conditional logistic regression, because the latter's likelihood is in general not equivalent to the case-crossover likelihood. This bias has been called *overlap bias* and can be substantial in the presence of strong temporal variability in event incidence and exposures. To quote Vines and Farrington (2001):

> It can be shown that for a dichotomous exposure, pairwise exchangeability is equivalent to requiring that the marginal probability of being exposed is the same in the case period and the control period. Thus if the exposure distribution is stationary over time, the Mantel-Haenszel estimator for the odds ratio is approximately asymptotically unbiased. When there are two or more control periods per case, global exchangeability (sufficient for the use of conditional logistic maximum likelihood estimator) is a stronger condition than pairwise exchangeability. In particular, stationarity of the exposure distribution over time is no longer sufficient. Over and above stationarity, it implies that the dependence in exposure between any two periods is identical. In matched case-control studies, exposures within each matched set can usually be assumed exchangeable, the numbering of controls being arbitrary. However this is not the case for case-crossover studies for which the time ordering of case and control periods is a fundamental aspect of the design. In most cases one might expect the exposure dependence between two periods to decline with the time interval between them. In practice, it is only safe to analyze case-crossover studies with M [referent times] >1 when the exposures across time periods are independent for each individual.

To appreciate this, consider the following example from Vines and Farrington (2001) examining the effect of a binary exposure (0 = non-exposed, 1 = exposed)

Table 8.8 Hypothetical case-crossover study from a cohort of 16,000 individuals illustrating how non-independence of exposure across time periods leads to biased odds ratio estimates from conditional logistic regression (Adapted from Vines and Farrington (2001))

Exposure pattern			Non-independence		Independence	
$t = 0$	$t = -1$	$t = -2$	Probability	Expected number[a]	Probability	Expected number
0	0	0	0	0	0.125	2
0	0	1	0	0	0.125	2
0	1	0	0	0	0.125	2
0	1	1	0.5	8	0.125	2
1	0	0	0.5	32	0.125	8
1	0	1	0	0	0.125	8
1	1	0	0	0	0.125	8
1	1	1	0	0	0.125	8

[a] Expected number = (population at risk) × (incidence proportion|case exposure at $t=0$) × (probability of exposure pattern)

on case occurrence in a case-crossover study of 16,000 individuals with two referent periods ($t=-2$ and $t=-1$) and the event period ($t=0$); in each period, the unconditional probability of exposure equals 0.5. The incidence proportion among the non-exposed is 0.001, and the incidence proportion among the exposed is 0.004, so the incidence odds ratio should approximately equal 4.0. Table 8.8 shows each combination of exposure in each of the three time periods and their respective probabilities under two distinct scenarios: non-independence and independence. Under non-independence, only two exposure patterns are possible, and exposure in both the referent times is dependent on (and opposite that of) exposure at $t=0$. A Mantel–Haenszel analysis of the data under non-independence yields the correct odds ratio of 4.0, while a conditional logistic regression analysis leads to a biased odds ratio of 6.6. In contrast, under independence all exposure patterns are equally likely to occur, so knowing exposure at $t=0$ provides no information on the distribution of exposures at $t=-1$ and $t=-2$. Both the Mantel–Haenszel and conditional logistic regression analyses yield an unbiased odds ratio of 4.0.

Non-independence might arise, for example, in a study of anti-inflammatory drug use in individuals allowed to self-medicate: during periods of elevated pain, people would be more likely to take the medication, while during periods of relief, they would be less likely. Because self-medication in one time interval is likely to provide some insight into self-medication in adjacent time intervals, the exposure information cannot be considered independent. When there is no true relationship between exposure and the outcome of interest, so long as exposure is stationary over time, the conditional logistic regression model should not erroneously suggest such a causal effect of exposure, although time of day, if explicitly modeled, could still demonstrate an effect on disease incidence. In contrast, the greater the true effect of exposure, the greater is the model's potential for bias, which can be either toward or away from the null.

8.4.4 Cohort Logic for Case-Crossover Studies

When time trends of exposure are present, the case-control logic of fixing the event time and comparing it to antecedent referent times leads to bias and must be discarded in favor of a counter-approach allowing the ordered referent time periods to be fixed, with the event time as random. This not only bears resemblance to the cohort paradigm and the self-controlled case-series method but also implies that referent times can be sampled both before and after the event time (i.e., ambidirectionally). Further innovations in case-crossover studies have exploited this design in studies of environmental exposures with transient effects exhibiting time trends. This is also consonant with the underlying structure for the case-crossover design as a proportional hazards model for a rare outcome for individual i with a constant baseline hazard who experiences the event at time t as

$$\lambda_i(t, x_{it}) = \lambda_i \exp(x_{it}\beta), \tag{8.12}$$

where x_{it} represents variables that are time dependent (Navidi 1998).

Two early examples by Navidi (1998) demonstrate the fallibility of unidirectional designs in the absence of exposure stationarity. The data in Table 8.9 show the experience of two cohorts over two successive time periods. The time trend in this example is obvious: one-third of people are non-exposed in Period 1, while two-thirds of people are non-exposed in Period 2. In both periods, the risk among the non-exposed is 0.001, the risk among the exposed is 0.002, and the risk ratio is 2. If a conventional unidirectional case-crossover study was performed with these data, where referent times anteceded fixed event times, it would only be possible to compare event times from Period 2 to referent times in Period 1. The single non-exposed case from sub-cohort 1 during Period 2 was exposed during Period 1, and the four exposed cases from sub-cohort 2 during Period 2 were non-exposed during Period 1. The ratio of these discordant event/referent pairs equals 4/1 = 4, which overestimates the true effect twofold. The bias arises because there were 2,000 individuals initially non-exposed in Period 1 that become exposed in Period 2, compared to only 1,000 individuals exposed in Period 1 who became non-exposed in Period 2. However, if the case-crossover study was instead bidirectional, then

Table 8.9 Hypothetical cohort study illustrating incidence of an outcome caused by exposure that varies over two time periods. The effects of exposure in each time period are independent (Adapted from Navidi (1998))

	Period 1				Period 2			
	Sub-cohort 1		Sub-cohort 2		Sub-cohort 1		Sub-cohort 2	
	Exposed	Non-exposed	Exposed	Non-exposed	Exposed	Non-exposed	Exposed	Non-exposed
Cases	2	0	0	2	0	1	4	0
Non-cases	1,000	0	0	2,000	0	999	1,996	0
Total	1,002	0	0	2,002	0	1,000	2,000	0

Table 8.10 Hypothetical cohort study illustrating incidence of an outcome caused by exposure in two time periods with differing baseline incidence. The effects of exposure in each time period are independent (Adapted from Navidi (1998))

	Period 1				Period 2			
	Sub-cohort 1		Sub-cohort 2		Sub-cohort 1		Sub-cohort 2	
	Exposed	Non-exposed	Exposed	Non-exposed	Exposed	Non-exposed	Exposed	Non-exposed
Cases	16	0	0	2	0	8	4	0
Non-cases	2,000	0	0	2,000	0	1,992	1,996	0
Total	2,016	0	0	2,002	0	2,000	2,000	0

the two exposed cases from sub-cohort 1 would have non-exposed referent times and the two non-exposed cases from sub-cohort 2 would have had exposed referent times. The matched pairs (Mantel–Haenszel) odds ratio would then become $(4+2)/(1+2) = 2$.

Table 8.10 shows an alternative example of how time trends in baseline risk, even in the absence of time trends in exposure, can lead to bias under a unidirectional referent sampling design. While there is no longer an association between the distribution of exposure among periods, as there was in Table 8.9, the risk in the non-exposed sub-cohort in Period 1 is 0.001, compared to the risk in the non-exposed sub-cohort in Period 2 of 0.004. With four exposed cases in Period 2 that were non-exposed in Period 1, and eight non-exposed cases in Period 2 that were exposed in Period 1, the ratio of these discordant event/referent pairs equals $4/8 = 0.5$. If a bidirectional design was instead employed, the 16 exposed cases that became non-exposed in Period 2 and the two non-exposed cases that became exposed in Period 2 would also be included, leading to a matched pairs odds ratio of $(4 + 16)/(8 + 2) = 2$, which equals the ratio of the exposed to non-exposed incidence in both sub-cohorts 1 and 2 ($0.008/0.004$ and $0.004/0.002$, respectively).

Although sampling referent times following the event is only strictly valid if an individual remains at risk following the event, and their exposures are unaffected by the event (i.e., exogenous, which is a reasonable assumption for environmental exposures), this bias is relatively minor compared to that incurred by only unidirectional sampling (Lumley and Levy 2000; Janes et al. 2005a) with rare events.

A greater understanding through likelihood specification of the assumptions inherent in cohort-centric sampling for case-crossover studies has been accompanied by a proliferation of newer designs. Navidi (1998) proposes what has come to be known as the full stratum bidirectional (FSBI) case-crossover design, although it is closely related to the self-controlled case-series design of Farrington (1995). This design utilizes the entire ordered time period of exposures as fixed, with the event time as random, leading referent times to be both before and after the event time. The underlying incidence rate is assumed to be constant over this period, which would not be correct in the face of seasonal or other period effects;

confounding by such temporal variation would need to be analytically controlled using either conditional logistic or Poisson regression.

Bateson and Schwartz (1999) propose taking referent times at identical intervals both before and after the event time to control for temporal confounding (e.g., by day of week). This has been referred to as the symmetric bidirectional (SBI) design. This approach fixes by design the referent times relative to the event time, sharing the case-control logic of unidirectional designs with the cohort logic of taking referent times before and after event times. Although taking referent times in close temporal proximity to the event time controls for long-term trends, this design requires exchangeability, without which conditional logistic regression estimates will be biased.

Lumley and Levy (2000) develop a referent sampling approach that requires the fixed ordered time period to be stratified by measures of time, such as month, week, or year, to control for these temporal effects, called the time-stratified (TS) design. The FSBI design can be considered a special case of the TS design with a single stratum encompassing all eligible referent times. Nevertheless, residual confounding from temporal effects may persist, leading to biased estimates (Whitaker et al. 2007). Analysis of this design can be achieved with either conditional logistic or Poisson regression, with time controlled with step functions using indicator variables. This model has been extended by Farrington and Whitaker (2006a) to allow for residual seasonal effects and called the time-season-stratified (TSS) design.

A modification of the SBI design is proposed by Navidi and Weinhandl (2002), called the semi-symmetric bidirectional (SSBI) design. Instead of two referent times symmetrically placed around the centered event time, only one referent time is randomly selected. This design assumes that temporal effects are constant, so referent times should be temporally in close proximity to the event time. Janes et al. (2005a) note that conditional logistic regression yields biased estimates unless an offset is incorporated into the model for days at the beginning and end of the exposure period.

This design is in turn modified by Janes et al. (2005a) to deal with event times at the respective ends of the exposure periods, where selecting only one referent time is possible. If a referent time was selected outside the exposure period, then the case would be eliminated from the study. This has been called the adjusted semi-symmetric bidirectional (ASSBI) design. Whitaker et al. (2007) note that this minor adjustment is a technical violation of the cohort logic of the case-crossover design because events at the two extreme times are excluded and that conditional logistic regression will yield valid estimates only so long as pairwise exchangeability is met.

In a review of the literature of 19 air pollution case-crossover studies between 1991 and 2004, Janes et al. (2005b) find that 63% utilized the SBI referent selection design, followed by 26% using the TS design. Nevertheless, these and other case-crossover designs all rely on assumptions that may be tenuous in practice. This leads Whitaker et al. (2007) to admonish epidemiologists against their use: "… we find little to recommend the continued use of the case-crossover approach in the analysis of environmental time series data: these methods are either biased or are special cases of more versatile methods. Time series regression is simple to use, does not suffer from overlap bias, and allows for flexible modeling of seasonality and time trends: this ought to be the method of choice."

8.5 Case-Time-Control Studies

One of the assumptions – and limitations – of the case-crossover design is that of stationarity (stability) of the distribution of study exposure over time. Such an assumption takes on a reasonable legitimacy for a transient exposure assessed over a relatively short referent time interval. However, as the period of prior exposure assessment lengthens, time trends in exposure from changing external factors may become emergent and evident, precluding the measurement of "usual frequency" of exposure. Period effects lead in turn to confounding.

Suissa (1995) recognizes time trends as a threat to validity in case-crossover studies, which leads him to propose a modification of a case-crossover study, which he calls a "case-time-control" study. The premise behind this design is that information about time trends in exposure can be obtained from individuals conventionally sampled as controls in a classical case-control study. This information can in turn be used to adjust the effect estimate from a case-crossover study, yielding less biased results by removing the confounding introduced by period effects.

To illustrate this bias practically, Suissa uses as a central example in pharmacoepidemiology (cf. chapter ▸Pharmacoepidemiology of this handbook) the problem of assessing drug effects in the face of confounding by indication. That is, prescription drug use is not only more likely among patients with more serious manifestations of an illness but as disease severity progresses over time, the therapeutic indication for such drug use is also likely to commensurately change as well. Because severity is typically both an independent predictor of a health outcome and an indication for treatment, nor is it an expected sequelae of therapeutic drug use, it fulfills the necessary criteria for confounding. Were severity an easily measurable host characteristic, it would be trivial to control for its confounding effects. However, severity within and between patients falls on a continuum that usually defies even imprecise measurement, such that even attempts to control for it would lead inevitably to meaningful residual confounding. Thus, neither conventional (i.e., case-control) nor case-only (e.g., case-crossover) studies would be capable of distinguishing drug (exposure) effects from temporal (confounding) changes in exposure.

Suissa's solution to this problem is to envision the odds of exposure (recognizing that sampling is based on outcome) in an individual as an exponential function of the following elements (retaining his notation):

$$\mathrm{odds}(E_{ijkl} = 1) = \exp(\mu + s_{il} + \pi_j + \vartheta_k), \tag{8.13}$$

where E represents a binary exposure, i represents the outcome group status (case or control), j represents a current or referent time period, k represents the outcome event in a particular period, and l represents the individual in group i; $\exp(\mu)$ is the overall exposure odds, $\exp(s_{il})$ is the participant-specific odds ratio, $\exp(\pi_j)$ is the period odds, and $\exp(\theta_k)$ is the odds corresponding to the outcome event. This function can be expressed separately for cases and controls and for the current and referent time periods. As will be seen shortly, this model is notable for its lack of interactions among the different elements. It can be shown that the effect measure of interest, $\exp(\theta_k)$, is not distinguishable from the nuisance period effect, $\exp(\pi_j)$, in the cases alone, but that the addition of a control series renders the former estimable.

This model, in which each individual's current and referent period constitute a matched pair, can be envisioned within the framework of conditional logistic regression. Again, utilizing Suissa's notation, let T denote the outcome variable which is the respective time period (1 = current, 0 = referent) for each participant's component of the matched pair, E the exposure, and G the outcome group status (case or control). The odds that the outcome T equals 1 is given by

$$\text{odds}(T = 1) = \exp(\beta_0 + \beta_1 E + \beta_2 (E \times G)). \tag{8.14}$$

The odds ratio corresponding to the effect of time period is $\exp(\beta_1)$, while the odds ratio corresponding to the effect of drug therapy on the outcome after removal of the period effect is $\exp(\beta_2)$.

The case-time-control study is not without its detractions. As Greenland (1996) points out, in addition to the usual assumptions of a case-crossover study (e.g., absence of carryover effects), the analysis can be confounded by the presence of unmeasured – and hence uncontrollable – confounders. Strictly speaking, for a case-time-control analysis to yield a valid point estimate of exposure, it must be assumed that there exist no interactions among any of the elements in (8.13). That is, the exposure-outcome association is unaffected by time period ($\pi_j \theta_k = 0$), the exposure-outcome association is unaffected by unmeasured confounders ($s_{il} \theta_k = 0$), and the exposure-time period association is unaffected by unmeasured confounders ($s_{il} \pi_j = 0$) (Suissa 1998). While the first assumption, that the effect measure remains stable, may be reasonable, particularly over shorter rather than longer time periods, the latter two assumptions are more problematic. The presence of unmeasured confounders is invariably a concern, though hardly unique to case-time-control studies. Those unmeasured variables may act on either the exposure effect as effect modifiers or on the time period effect as confounders.

The utility of this method, then, rests on the veracity of unverifiable assumptions of no-confounding by indicators of disease severity. A more recent modification of this approach, called the "case-case-time-control" design (Wang et al. 2011) relies solely on future cases as the source of controls. Greenland (1996) points out the problem of selecting between competing study designs to cope with the problem of confounding by indication when it is unclear which design would yield a less biased result. Without a recommendation that uniformly applies to all study settings, it may ultimately be advisable when possible to employ a sensitivity analysis of two or more designs within the same cohort in order to see how their different properties and assumptions can affect conclusions.

8.6 Self-Controlled Case-Series Methods

Considerable interest in the last decade has focused on the putative relationship between vaccines and adverse outcomes (both acute and chronic) that are too rare to be detected in limited-size safety trials. If such outcomes are sequelae to vaccines, they are unlikely to be recognized until substantial utilization and experience in populations have been realized. Conventional non-experimental studies suffer from methodological inadequacies: depending on the rarity of the event, cohort studies would likely be prohibitively large, requiring populations difficult to enumerate, and hospital-based case-control studies would be compromised by an unclear choice of a source population and valid control series, and the issue of transient duration of vaccine effect. The relative rarity of cases would likely also compromise the investigator's ability to control for the myriad of factors known to be related to choice and timing of vaccination.

One design option to address this challenge of studying time-variant exposures with transient effects is proposed by Farrington (1995). The study population is comprised of a case series of individuals belonging to an underlying source population who experience one or more instances of the outcome of interest during a defined observation period and who have retrospectively obtained vaccination (exposure) histories. Although it only includes cases, its design allows the self-control of within-individual time-invariant confounders (e.g., genetic propensity and environmental factors) and coupled with outcome rarity allows the estimation of consistent effect measures of relative incidence. The design differs from the case-crossover model, which assumes stationarity of the exposure distribution, and allows the assessment of age and time effects, as well as multiple events (Whitaker et al. 2006). In addition, when exposure is not rare, this design is nearly as efficient as a study of the entire underlying cohort (Farrington 2004).

A number of key assumptions are inherent in this model (Whitaker et al. 2009). While exposure is time dependent, in the case of recurrent events, it is assumed that one event does not affect the probability of subsequent exposure, although the authors have proposed modifications if there is either one or multiple exposures by excluding all cases whose exposures were not prior to the event. A second assumption is that the duration of the observation period must be independent of the occurrence of one or more events.

The model (in the notation of Whitaker et al. 2009) can be represented by the following:

$$\lambda_{ijk} = \exp(\varphi_i + \alpha_j + \beta_k), \tag{8.15}$$

where λ_{ijk} is the incidence rate of the ith individual in the jth age stratum and kth risk period, φ_i is the effect of individual i, α_j is the effect of age group j, and β_k is the effect of exposure in risk period k. n_{ijk} is the number of events occurring within a risk period of length e_{ijk}, which are assumed to arise in a non-homogeneous, age-dependent Poisson process, that is, $n_{ijk} \sim$ Poisson $(\lambda_{ijk}e_{ijk})$. Non-homogeneity reflects that incidence changes within the observation period due to periods of intermittent altered and attenuated risk from exposure, and agedependent reflects that incidence is allowed to vary across age strata while homogeneous within age strata. Exp(β_k) is the incidence rate ratio comparing the incidence in the kth risk period relative to the control period (i.e., $k = 0$). Medical knowledge informs the definition of the risk periods coming before or after exposure k, where $k = 1, 2, \ldots$, K, during which cases are at presumptive lower or higher risk of an event. Any other times within the observation period before, after, or between the risk periods constitute the control periods, indexed by $k = 0$. Cases will thus contribute time within some (but not necessarily all) combinations of age and risk periods. Fixed covariates can be included in the model if there is suspicion that they may modify the effect of exposure on event incidence by including interaction terms.

The approach is flexible enough to accommodate multiple risk periods, as when multivalent vaccines are evaluated in which the respective components have different periods of increased risk. It can straightforwardly accommodate multiple events, and if the events cluster together in time, but the clusters themselves are longitudinally independent, then it is possible to use the first event in each cluster of events while still using multiple clusters.

Care must be exercised in the construct of a case series to assess effects of transient exposures, whether they are vaccines or other pharmacological agents. The probability of case selection should be statistically independent of exposure status; this should be achieved if a census of cases conditional on eligibility criteria is taken or if exposure information is retrieved only after sampled cases are enrolled. Ages and observation period studied should be selected to achieve periods of time both in and out of risk periods. The choice of risk period is non-trivial and must be based on prior medical knowledge and biological plausibility. A single risk period can also be subdivided into non-overlapping intervals to detect when, within the period, exposure effects are most pronounced. Although the model was designed to model age interval effects, other time-dependent categorical variables can be substituted or added (e.g., season).

Figure 8.3 conceptually demonstrates how vaccines administered to children in two age groups experiencing immune-mediated hemolytic anemia could be evaluated. Each child's 2-year period of observation ($j = 1, 2$) is subdivided into four risk periods: baseline ($k = 0$) and three successive 2-week intervals ($k = 1, 2, 3$) in which vaccines could potentially trigger the immune system to begin hemolysis. Vaccines were administered on the first day of the first risk period ($k = 1$) under the premise that an acute hypersensitivity reaction could lead to an immediate hemolytic crisis. If this is not medically plausible in the context of other problems, then a lag period of no risk ($k = 0$) can be inserted between the day of vaccination and the first day of the first risk period. Table 8.11 translates these patients' experiences into the data model appropriate for analysis (e.g., using conditional Poisson regression). Although receiving a vaccination in the observation period guarantees that patients will contribute time to all levels of k (unless patients receive vaccines within 6 weeks of the end of the observation period), not all combinations of age and risk categories need have intervals populated with time (e.g., patient three).

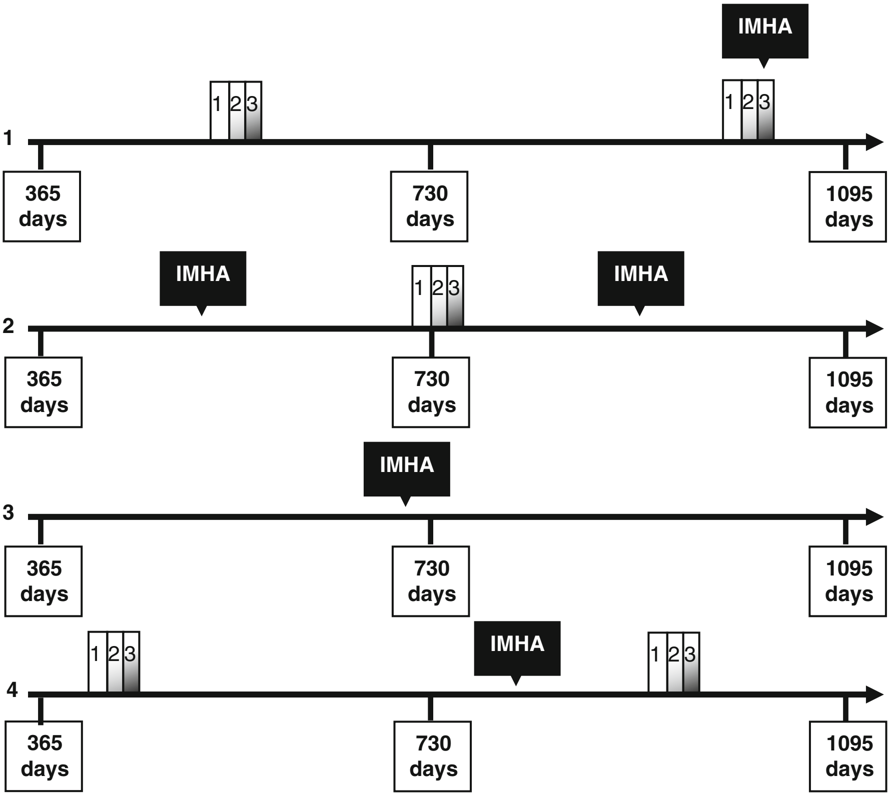


Fig. 8.3 Four hypothetical cases of immune-mediated hemolytic anemia (IMHA) under observation from 1 to 3 years of age. *Black box* arrows indicate time of onset of IMHA. Vaccine effects were postulated to occur in three successive risk periods of 14 days duration (*vertical boxes* numbered 1, 2, and 3). Vaccination was given on the first day of the first risk period (k = 1), so no lag period of no risk was assumed

The self-controlled case-series methods continue to evolve. Farrington and Whitaker (2006b) modify the approach to modeling baseline ($k = 0$) incidence by using a semi-parametric approach. When the assumption that the event does not affect the probability of subsequent exposure (e.g., a person dies from the event), then adaptations can be made to estimate hypothetical time at risk. The method has also been adapted for use in prospective post-marketing pharmacosurveillance for periods of short risk as well as long and indefinite risk periods. These methods are discussed further in Whitaker et al. (2009).

Sample size formulas for self-controlled case series have been developed by Musonda et al. (2006). A useful website has been created to update users on the methodology of self-controlled case series, as well as to provide instruction and guidance to creating datasets partitioned by age, risk period, and potentially other covariates, and to performing regression analyses using commercial software packages: http://statistics.open.ac.uk/sccs (accessed on November 28, 2011).

Table 8.11 Data layout for four patients (ages 1–3 years) in Fig. 8.3 with one or more episodes of immune-mediated hemolytic anemia. Vaccine effects were postulated to occur in three successive intervals of 14 days duration. Interval lengths of 0 days (e.g., where no periods of vaccine risk occurred in an age group) are not shown

Patient	Age group	Period of vaccine risk (K)	Number of events (n_{ijk})	Interval length (e_{ijk})
1	365–729	0	0	180
1	365–729	1	0	14
1	365–729	2	0	14
1	365–729	3	0	14
1	365–729	0	0	143
1	730–1,094	0	0	270
1	730–1,094	1	0	14
1	730–1,094	2	0	14
1	730–1,094	3	1	14
1	730–1,094	0	0	53
2	365–729	0	1	351
2	365–729	1	0	14
2	730–1,094	2	0	14
2	730–1,094	3	0	14
2	730–1,094	0	1	337
3	365–729	0	1	365
3	730–1,094	0	0	365
4	365–729	0	0	70
4	365–729	1	0	14
4	365–729	2	0	14
4	365–729	3	0	14
4	365–729	0	0	253
4	730–1,094	0	1	175
4	730–1,094	1	0	14
4	730–1,094	2	0	14
4	730–1,094	3	0	14
4	730–1,094	0	0	148

8.7 Case-Specular Studies

Earlier discussions surrounding the study of exposures with intermittent effects and whose durations are transient have promoted a counterfactual paradigm as superior in conception and function to a study base paradigm in motivating the selection of a comparison series for cases experiencing a health outcome. This builds upon a conceptual framework that envisions the measurement of a unit-level effect as a causal contrast between an individual's outcome under one condition and what that individual's outcome would have potentially been under the absence of that condition. This contrast has been extended to population level effects as well: the contrast between the incidence of an outcome in an assemblage of individuals sharing some exposure and what the incidence of the outcome potentially would have been had the exposure been absent or its effect annulled.

With little effort, these principles can be extended to population studies in which sampling is based on outcome, rather than exposure. In an actual case-control study, antecedent exposure histories are compared with those from a control series drawn from a source population at risk. This can be juxtaposed with a "potential exposures" vision: controls should ideally reflect the potential distribution of exposure in the case series had any causal link been abolished prior to the outcome initiation.

In the context of the case-crossover study, this concept has a natural correspondence: comparing cases' proximate (relative to the incident outcome event) exposure history to the same cases' potential proximate exposure history in the absence of causation. Confounding by factors that are inherent time-invariant properties of the individual (e.g., genetic predisposition) becomes irrelevant because each individual's actual exposure, together with her/his counterfactual exposure, is analyzed as if from a matched pair in a unique stratum.

By adapting such reasoning to other settings, novel study design approaches become both possible and practical. One example is the residential case-specular method of studying the hypothesized causal relationship between wire codes (used as an indirect estimate of historical exposure to electromagnetic fields) and childhood cancers (Zaffanella et al. 1998). Wertheimer and Leeper (1979) initially promote the use of power line wire code categories, which are functions of wire thickness and distance from power lines to residences (Table 8.9). The use of wire codes as an imperfect surrogate for field measurements has a practical implication: household members need not participate for exposure measurements to be taken, which eliminates selection bias if voluntary enrollment is associated with wire code and allows for larger study population sizes (Ebi et al. 1999).

Unfortunately, use of wire codes has an important drawback: spatial proximity of power lines to residences is not only a proxy for exposure to electromagnetic fields but also can serve as a proxy for other characteristics, either unknown, unmeasured, or unmeasurable characteristics of neighborhoods, such as socioeconomic status, traffic congestion and air pollution, and environmental contamination. These may marginally or jointly have effects on health issues such as childhood leukemia. Therefore, any residential study of the electromagnetic field hypothesis would necessitate distinguishing two distinct, but not necessarily competing, effects: that of the electromagnetic fields and those other effects that are intrinsic characteristics of the neighborhood.

The case-specular design was conceived to mitigate the problem of distinguishing electromagnetic field effects, if any, from others that defied analytical control. The premise is that case residences with quantifiable wire code exposures could be contrasted with the potential wire code exposures of purely hypothetical (counterfactual) residences. In this design inception, the counterfactual residence is specular: an imaginary or virtual residence created by symmetrically (sagitally) reflecting a case residence across the center of its street, creating a matched case-specular pair (Fig. 8.4). Matched analyses of case-specular data with wire code exposures, with matching on residential street, should then distinguish potential

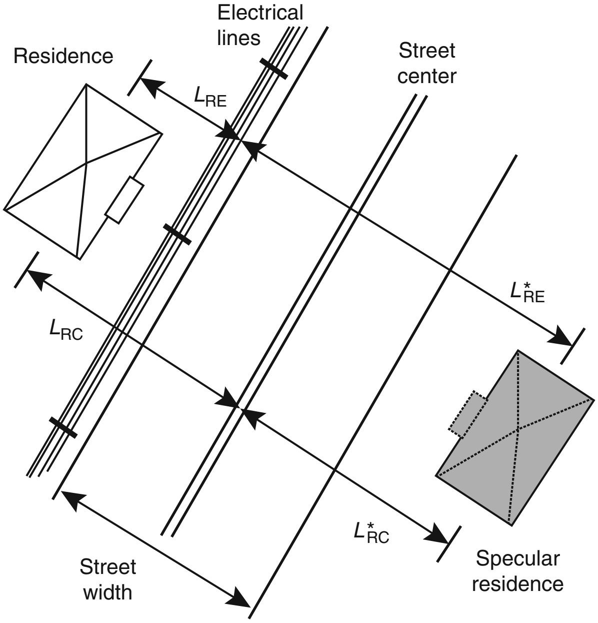


Fig. 8.4 Example of construction of a specular residence on a street with case residence on the same street side as electrical lines. Note that the distance from the residence to the center of the street, L_{RC}, is equal to the distance from the center of the street to the specular residence, L^*_{RC}. Also, L_{RE} = distance from the case residence to the electrical lines, while L^*_{RE} = distance from the specular residence to the electrical lines

effects of electromagnetic frequencies from the neighborhood "effect." Alternative speculars have also been proposed, including reflected images of power lines and translational or rotational movement of houses or power grids (Allard 1998).

Following the approach of Zaffanella et al. (1998), the distance from a case residence to the center of the street (L_{RC}) is used to create a specular control residence with an identical, albeit hypothetical, distance (L^*_{RC}). Two measurements are then taken to assist with determining exposure: the distance from the case residence to the electrical lines (L_{RE}) and what the distance from the specular control residence would be, had it existed, to the electrical lines (L^*_{RE}). The wire code category of the case residence (Table 8.12) can then be contrasted with the counterfactual wire code of the specular residence. The matching of this pair of residences is so spatially fine that, apart from potential discordance in wire code, it is plausible to assume that most if not all other environmental or social determinants of outcome are concordant within the matched pairs. If the wire code acts only as a surrogate for such neighborhood risk factors, then no residual association should exist between the higher current wire codes postulated to be related to cancer risk and case residence. The statistical analysis for this design is typical of those

Table 8.12 Wire code ordinal categories as described by Wertheimer and Leeper (1979) and Ebi et al. (1999)

Category[a]	Line class[b]	Distance to residence
UG	Not applicable	All distribution lines within 46 m (150 feet) are located underground
VLCC	6	End-pole situations
OLCC	5	Within 46 m (150 feet)
	4	Within 46 m (150 feet)
	1–3	From the OHCC upper distance to 46 m (150 feet)
OHCC	3	Within 15 m (50 feet)
	2	Between 7.5 and 19.5 m (25–64 feet)
	1	Between 15 and 39.5 m (50–129 feet)
VHCC	2	Within 7.5 m (25 feet)
	1	Within 15 m (50 feet)

[a] *VHCC* very high current configurations, *OHCC* ordinary high current configurations, *OLCC* ordinary low current configurations, *VLCC* very low current configuration, *UG* underground wiring

[b] Class 1 = transmission lines, three-phase primary distribution lines with thick wires or with multiple circuits; Class 2 = three-phase primary distribution lines with thin wires; Class 3 = long first span secondary distribution lines; Class 4 = second span secondary distribution lines; Class 5 = short first span secondary distribution lines; Class 6 = end-pole situations

for matched pair data (e.g., conditional logistic regression), with the polytomous exposures corresponding to the wire code categories.

Zaffanella et al. (1998) note, in contrast, that if the electromagnetic field hypothesis is correct, a preponderance of case households should occur on the same side of a street as power lines and that case residences should have, on average, higher wire codes than specular residences. Several assumptions are necessary for these predictions, if empirically found in a study, to have causal interpretation. The first, called "symmetry of the residence-specular probability matrix," implies probabilistic independence between placement of power lines and placement of residences on the sides of a street. The second assumption is an implicit "randomization" of residences on both sides of a street with respect to unmeasured, unmeasurable, or unknown confounders (analogous to the customary no-confounding assumption underlying causal interpretation of all non-experimental studies). The third assumption is no systematic misclassification of wire codes by residence type (case or specular), a problem that cannot be mitigated by blinding if only case residences are studied in situ. Differential misclassification may be a particular problem due to subjectivity in assigning wire codes to specular residences.

The case-specular method shares an advantage of neighbor-matched case-control studies: the control of confounding attributable to intrinsic properties of the neighborhood. It also enjoys an economical advantage of not requiring environmental measurements from a control group. It has the disadvantages of requiring the specification of speculars in unrealistic situations, high frequencies of concordant residence-specular pairs (particularly when power lines are located behind homes and uniform wire classes are used), and the inability to verify the assumption of symmetry of the residence-specular probability matrix without a validation control series. For further issues and caveats related to the analysis of case-specular and other forms of case-only (case-distribution) studies, and the incorporation of controls into the study (i.e., case-control-specular), see Greenland (1999).

8.8 Conclusions

The designs addressed in this chapter, all of relatively recent incarnation compared to their progenitor observational counterparts, should be appreciated less for their differences and more for their common lineage. These more recent designs are the evolutionary culmination of a long-held view that causal inference in populations is not fundamentally a comparison of individuals to each other, but is instead a collective comparison of single individuals to themselves. This view is rendered practical via the use of empirically derived or hypothetical exposure distributions, such as in case-crossover and case-specular studies, respectively, thus supplanting the need for an external comparison group. It seems inevitable that future advances in design, analysis, and efficiency will build upon this conceptualization. As epidemiologists employ these new designs and contrast them with older ones, the designs will undoubtedly undergo even greater scrutiny with respect to their assumptions and limitations. Such circumspection will particularly be necessary when conflicts in findings resulting from use of the different designs arise. These conflicts, however, should be regarded as the natural consequence of a progressive series of improvements in epidemiological methods that will inevitably lead to more valid assessment of potential causal relations.

Acknowledgements I would like to gratefully acknowledge the invaluable advice of the anonymous reviewer of the first edition of this chapter, who suggested numerous additional references and topics deserving greater attention, in turn helping to greatly improve its second edition.

References

Allard R (1998) The residential case-specular method to study wire codes, magnetic fields, and disease. Epidemiology 9:475–476

Barlow WE, Ichikawa L, Rosner D, Izumi S (1999) Analysis of case-cohort designs. J Clin Epidemiol 52:1165–1172

Bateson TF, Schwartz J (1999) Control for seasonal variation and time trend in case-crossover studies of acute effects of environmental exposures. Epidemiology 10:539–544

Breslow NE, Chatterjee N (1999) Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. Appl Stat 48:457–468

Breslow N, Day NE (1987) Statistical methods in cancer research, volume II: The design and analysis of cohort studies. International Agency for Research on Cancer, Lyon

Breslow NE, Lubin H, Marek P, Langholz B (1983) Multiplicative models and cohort analysis. J Am Stat Assoc 78:1–12

Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M (2009a) Using the whole cohort in the analysis of case-cohort data. Am J Epidemiol 169:1398–1405

Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M (2009b) Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. Stat Biosci 1:32

Cornfield J (1951) A method of estimating comparative rates from clinical data. J Natl Cancer Inst 11:1269–1275

D'Angio GJ, Breslow N, Beckwith B, Evans A, Baum E, Delorimier A, Fernbach D, Hrabovsky E, Jones B, Kelalis P, Othersen B, Teft M, Thomas PRM (1989) Treatment of Wilms' tumor. Cancer 64:349–360

Deville JC, Särndal CE (1992) Calibration estimator in survey sampling. J Am Stat Assoc 87: 376–382

Ebi KL, Zaffanella LE, Greenland S (1999) Application of the case-specular method to two studies of wire codes and childhood cancers. Epidemiology 10:398–404

Farrington CP (1995) Relative incidence estimation from case series for vaccine safety evaluation. Biometrics 31:228–235

Farrington CP (2004) Control without separate controls: evaluation of vaccine safety using case-only methods. Vaccine 22:2064–2070

Farrington CP, Whitaker HJ (2006a) Semiparametric analysis of case series data. J R Stat Soc 55:553–594

Farrington CP, Whitaker HJ (2006b) Semiparametric analysis of case series data. Appl Stat 55: 553–594

Greenland S (1996) Confounding and exposure trends in case-crossover and case-time-control designs. Epidemiology 7:231–239

Greenland S (1999) A unified approach to the analysis of case-distribution (case-only) studies. Stat Med 18:1–15

Greenland S, Robins JM (1985) Estimation of a common effect parameter from sparse follow-up data. Biometrics 41:55–68

Greenland S, Thomas DC (1982) On the need for the rare disease assumption in case-control studies. Am J Epidemiol 116:547–553

Greenland S, Thomas DC, Morgenstern H (1986) The rare-disease assumption revisited. A critique of "estimators of relative risk for case-control studies". Am J Epidemiol 124:869–883

Janes H, Sheppard L, Lumley T (2005a) Case-crossover analyses of air pollution exposure data: referent selection strategies and their implications for bias. Epidemiology 16:717–726

Janes H, Sheppard L, Lumley T (2005b) Overlap bias in the case-crossover design, with application to air pollution exposures. Stat Med 24:285–300

Kulich M, Lin DY (2004) Improving the efficiency of relative-risk estimation in case-cohort studies. J Am Stat Assoc 99:832–844

Kupper LL, McMichael AJ, Spirtas R (1975) A hybrid epidemiologic study design useful for estimating relative risk. J Am Stat Assoc 70:524–528

L'Abbé KA, Howe GR, Burch JD, Miller AB, Abbatt J, Band P, Choi W, Du J, Feather J, Gallagher R, Hill G, Matthews V (1991) Radon exposure, cigarette smoking, and other mining experience in the Beaverlodge uranium miners cohort. Health Phys 60:489–495

Langholz B (2005) Counter-matching. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, 2nd edn. Wiley, Chichester

Langholz B, Borgan Ø (1995) Counter-matching: a stratified nested case-control sampling method. Biometrika 82:69–79

Langholz B, Clayton D (1994) Sampling strategies in nested case-control studies. Environ Health Perspect 102(Suppl 8):47–51

Langholz B, Thomas DC (1990) Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. Am J Epidemiol 131:169–176

Lin DY, Ying Z (1993) Cox regression with incomplete covariate measurements. J Am Stat Assoc 88:1341–1349

Little RJ, Rubin DR (2000) Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. Annu Rev Public Health 21:121–145

Lumley T (2008) Analysis of complex samples in R. Surv Stat 57:20–25

Lumley T, Levy D (2000) Bias in the case-crossover design: implications for studies of air pollution. Environmetrics 11:689–704

Maclure M (1991) The case-crossover design: a method for studying transient effects on the risk of acute events. Am J Epidemiol 133:144–153

Mark SD, Katki HA (2006) Specifying and implementing nonparametric and semiparametric survival estimators in two-stage (sampled) cohort studies with missing case data. J Am Stat Assoc 101:460–471

Miettinen OS (1976) Estimability and estimation in case-referent studies. Am J Epidemiol 103:226–235

Miettinen OS (1982) Design options in epidemiologic research: an update. Scand J Work Environ Health 8(Suppl 1):7–14

Mittleman MA, Maclure M, Robins JM (1995) Control sampling strategies for case-crossover studies: an assessment of relative efficiency. Am J Epidemiol 142:91–98

Musonda P, Farrington CP, Whitaker HJ (2006) Sample sizes for self-controlled case series studies. Stat Med 25:2618–2631. Erratum in: Statistics in Medicine 2008 27:4854–4855

Navidi W (1998) Bidirectional case-crossover designs for exposures with time trends. Biometrics 54:596–605

Navidi W, Weinhandl E (2002) Risk set sampling for case-crossover designs. Epidemiology 13:100–105

Prentice RL (1986) A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 73:1–11

Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 89:846–866

Sato T (1992a) Maximum likelihood estimation of the risk ratio in case-cohort studies. Biometrics 48:1215–1221

Sato T (1992b) Estimation of a common risk ratio in stratified case-cohort studies. Stat Med 11:1599–1605

Sato T (1994) Risk ratio estimation in case-cohort studies. Environ Health Perspect 102(Suppl 8):53–56

Steenland K, Deddens JA (1997) Increased precision using countermatching in nested case-control studies. Epidemiology 8:238–242

Suissa S (1995) The case-time-control study. Epidemiology 6:248–253

Suissa S (1998) The case-time-control design: further assumptions and conditions. Epidemiology 9:441–445

Suissa S, Dell'Aniello S, Martinez C (2010) The multitime case-control design for time-varying exposures. Epidemiology 21:876–883

The Atherosclerosis Risk in Communities (ARIC) study: design and objectives (1989) The ARIC Investigators. Am J Epidemiol 129:687–702

Vines SK, Farrington CP (2001) Within-subject exposure dependency in case-crossover studies. Stat Med 20:3039–3049

Wacholder S (1991) Practical considerations in choosing between case-cohort and nested case-control designs. Epidemiology 2:155–158

Wacholder S, Boivin JF (1987) External comparisons with the case-cohort design. Am J Epidemiol 126:1198–1209

Walker AM (1982) Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known. Biometrics 38:1025–1032

Wang S, Linkletter C, Maclure M, Dore D, Mor V, Buka S, Wellenius GA (2011) Future cases as present controls to adjust for exposure trend bias in case-only studies. Epidemiology 22: 568–574

Wertheimer N, Leeper E (1979) Electrical wiring configurations and childhood cancer. Am J Epidemiol 109:273–284

Whitaker HJ, Farrington CP, Spiessens B, Musonda P (2006) Tutorial in biostatistics: the self-controlled case series method. Stat Med 25:1768–1797s

Whitaker HJ, Hocine MN, Farrington CP (2007) On case-crossover methods for environmental time series data. Environmetrics 18:157–171

Whitaker HJ, Hocine MN, Farrington CP (2009) The methodology of self-controlled case series studies. Stat Methods Med Res 18:7–26

White JE (1982) A two stage design for the study of the relationship between a rare exposure and a rare disease. Ame J Epidemiol 115:119–128

Witt MB (2009) Overview of software that will produce sample weight adjustments. Joint statistical meetings, section on survey research methods. http://www.amstat.org/sections/srms/Proceedings/y2009/Files/304290.pdf. Accessed 28 Nov 2011

Zaffanella LE, Savitz DA, Greenland S, Ebi KL (1998) The residential case-specular method to study wire codes, magnetic fields, and disease. Epidemiology 9:16–20

Zhang H, Schaubel DE, Kalbfleisch JD (2011) Proportional hazards regression for the analysis of clustered survival data from case-cohort studies. Biometrics 67:18–28

Intervention Trials

9

Silvia Franceschi and Martyn Plummer

Contents

S. Franceschi (✉) • M. Plummer
Infections and Cancer Epidemiology Group, International Agency for Research on Cancer, Lyon, France

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_9,

9.1 Introduction

Most causes of premature death are avoidable. This is true not only of infectious diseases, but also a large proportion of cardiovascular diseases and cancer (Doll and Peto 1981). A major goal of public health is to find the causes of disease in a population and then intervene to remove them.

As discussed by Doll (2002), an agent may be considered to be a cause of a disease if increased exposure to the agent is followed by an increased risk of the disease and decreased exposure by a decreased risk. This is an empirical definition which may be tested without reference to a specific mechanism. It is particularly useful for chronic diseases, such as cancer, which may take decades to develop and in which the disease process may involve a variety of preclinical changes before the disease manifests itself. Intervention to remove a cause of disease may then range from behavioral changes, such as tobacco control (IARC 2007), to minimizing consequences of accumulated damages by, for example, regression of precancerous lesions using antioxidant vitamins (Stewart et al. 1996).

Descriptive and observational epidemiological studies have provided considerable evidence for causal relationships and have, in some instances, provided the final answer (Doll 2002). However, observational studies are not always sufficient to motivate large-scale public health interventions as they have some important limitations. Firstly, when relative risks of the order 2 or less are observed, it is difficult to rule out bias and confounding as possible explanations for the association. A second limitation of observational studies is that they rely on "experiments of nature" – unplanned variation in exposure within and between populations – and cannot therefore evaluate the effect of interventions that attempt to block the disease process in a way that is not found in nature. Two examples are a cholesterol-lowering drug (Heart Protection Study Collaborative Group 2002b) or a prophylactic vaccine against human papilloma virus (HPV) (Koutsky et al. 2002). Both of these interventions are motivated by a large body of observational evidence as well as understanding of the mechanism of disease. However, the magnitude of the benefit from intervention cannot be evaluated from observational data. The third limitation of observational studies is that it is very difficult to balance the benefits of intervention against possible risks. Finally, observational data do not always provide evidence that exposure to an agent preceded incidence, which is a requirement for establishing causality. All of these limitations can be overcome by a properly conducted intervention trial. An intervention trial is an experiment to evaluate the efficacy of an intervention so that its more widespread use can be justified.

Intervention trials ideally take the form of a randomized controlled trial (RCT) in which the intervention is compared with a control and the allocation to treatment or control is randomized. The control may be a placebo or an alternative intervention. RCTs are often used for *therapeutic trials* in which different treatments for a given disease are compared in a clinical setting. This application of RCTs is beyond the scope of this chapter, which is concerned with trials on healthy or apparently healthy individuals with the aim of preventing future morbidity or mortality. We refer to such trials as *preventive trials*.

Since RCTs offer a unique opportunity to eliminate the problems that beset observational studies, the results of such trials are generally considered to be a "gold standard," and are often taken to outweigh previous observational evidence in the case of discordant results. However, the advantages of RCTs are easily lost through poor conduct or analysis. Hence, RCTs are held to much higher standards of conduct and reporting (see Sect. 9.1.1).

9.1.1 Guidelines for Reporting of Clinical Trials: The CONSORT Statement

The CONSORT (Consolidated Standards of Reporting Trials) statement is a set of guidelines for reporting of RCTs (http://www.consort-statement.org). These guidelines take the form of a checklist of 22 items (see Table 9.1) which should be included in a report of a randomized trial and a model flow chart to show the flow of participants through the trial (see Fig. 9.1).

The guidelines were created by an international group of clinical trialists, statisticians, epidemiologists, and biomedical journals in an effort to improve the quality of reporting of RCTs, which several reviews had shown to be inadequate. Without adequate reporting, it is not possible for a reader to judge the quality of a trial and so trust its conclusions.

The original CONSORT statement was published in 1996 (Begg et al. 1996). It was revised in 2001 and is available in a short form (Moher et al. 2001) and a long form with explanation and elaboration (Altman et al. 2001). The guidelines have been adopted by a growing number of biomedical journals and editorial committees.

9.2 Therapeutic Versus Preventive Trials

The methodological considerations for therapeutic and preventive trials are very similar. In particular, in the last two decades, therapeutic trials have grown in size. Some therapeutic trials have been able to randomize many thousands of individuals in studies of breast cancer (Early Breast Cancer Trialists' Collaborative Group 1998), heart disease (ISIS-2 (Second International Study of Infarct Survival) Collaborative Group 1988), and stroke (CAST (Chinese Acute Stroke Trial) Collaborative Group 1997). In this chapter, we will rely heavily upon the experience gained in the performance of therapeutic trials. (For details on prevention in general, please refer to chapter ►Screening of this handbook.)

Although there is no clear dividing line between therapeutic and preventive trials, a crucial factor distinguishing the two kinds of trial is the "exposure window." In a therapeutic trial, the timing of an intervention is determined by a disease indication. In a prevention trial, the intervention may take place at any point along a disease process that may last several decades. The appropriate timing of the intervention is not always evident. Often, the decision on when to intervene is

Table 9.1 Checklist of items to include when reporting a randomized trial (Source: http://www.consort-statement.org)

Paper section and topic	Item	Description
TITLE AND ABSTRACT	1	*How participants were allocated to interventions* (e.g., "random allocation," "randomized," or "randomly assigned")
INTRODUCTION		
Background	2	*Scientific background and explanation of rationale*
METHODS		
Participants	3	*Eligibility criteria for participants and the settings and locations where the data were collected*
Interventions	4	*Precise details of the interventions intended for each group and how and when they were actually administered*
Objectives	5	*Specific objectives and hypotheses*
Outcomes	6	*Clearly defined primary and secondary outcome measures* and, when applicable, any *methods used to enhance the quality of measurements* (e.g., multiple observations, training of assessors)
Sample size	7	*How sample size was determined* and, when applicable, *explanation of any interim analyses and stopping rules*
Randomization – sequence generation	8	*Method used to generate the random allocation sequence*, including *details of any restriction* (e.g., blocking, stratification)
Randomization – allocation concealment	9	*Method used to implement the random allocation sequence* (e.g., numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned
Randomization – implementation	10	*Who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups*
Binding (masking)	11	*Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment.* When relevant, *how the success of blinding was evaluated*
Statistical methods	12	*Statistical methods used to compare groups for primary outcome(s); methods for additional analyses*, such as subgroup analyses and adjusted analyses
RESULTS		
Participant flow	13	*Flow of participants through each stage* (a diagram is strongly recommended). Specifically, for each group report the numbers of participants randomly assigned, receiving intended treatment, completing the study protocol, and analyzed for the primary outcome. *Describe protocol deviations from study as planned, together with reasons*

(continued)

Table 9.1 (Continued)

Paper section and topic	Item	Description
Recruitment	14	*Dates defining the periods of recruitment and follow-up*
Baseline data	15	*Baseline demographic and clinical characteristics of each group*
Numbers analyzed	16	*Number of participants (denominator) in each group included in each analysis and whether the analysis was by "intention-to-treat."* State the results in absolute numbers when feasible (e.g., 10/20, not 50%)
Outcomes and estimation	17	*For each primary and secondary outcome, a summary of results for each group, and the estimated effect size and its precision* (e.g., 95% confidence interval)
Ancillary analyses	18	*Address multiplicity by reporting any other analyses performed*, including subgroup analyses and adjusted analyses, indicating those prespecified and those exploratory
Adverse events	19	*All important adverse events or side effects in each intervention group*
DISCUSSION		
Interpretation	20	*Interpretation of the results*, taking into account study hypotheses, sources of potential bias or imprecision, and the dangers associated with multiplicity of analyses and outcomes
Generalizability	21	*Generalizability (external validity) of the trial findings*
Overall evidence	22	*General interpretation of the results in the context of current evidence*

determined by the practical requirement to have a sufficient number of cases of disease in the control group by the end of the trial. This limits interventions to individuals who are at high risk, either because they have long-term exposure to known risk factors or because they have advanced preclinical disease. Examples of such trials are the prevention of recurrence of colorectal adenomas (Bonithon-Kopp et al. 2000; Jacobs et al. 2002) and prevention of lung cancer in middle-aged heavy smokers or workers exposed to asbestos (Omenn et al. 1996a; ATBC (Alpha-Tocopherol Beta Carotene) Cancer Prevention Study Group 1994). The paradox of such late interventions, however, is that they may miss the "exposure window" within which the intervention may possibly have a protective effect.

9.2.1 Surrogate Endpoints

It is difficult to conduct intervention trials using disease endpoints, such as cancer, that are rare or have a long latency. Such studies may require very large numbers,

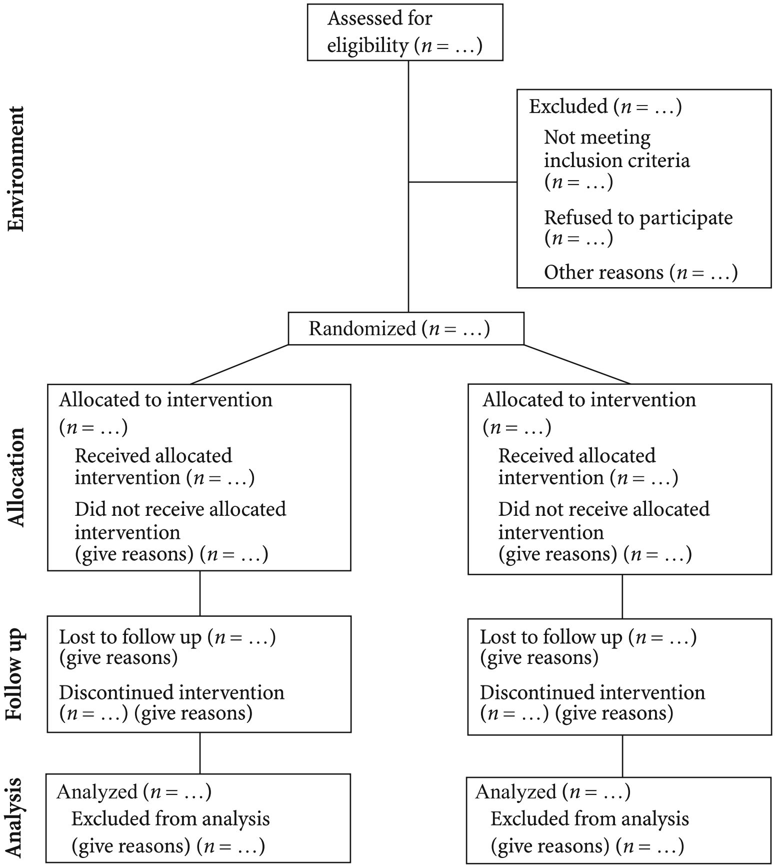

Assessed for eligibility (n = …)
Excluded (n = …)
Not meeting inclusion criteria (n = …)
Refused to participate (n = …)
Other reasons (n = …)
Randomized (n = …)
Allocated to intervention (n = …)
Received allocated intervention (n = …)
Did not receive allocated intervention (give reasons) (n = …)
Allocated to intervention (n = …)
Received allocated intervention (n = …)
Did not receive allocated intervention (give reasons) (n = …)
Lost to follow up (n = …) (give reasons)
Discontinued intervention (n = …) (give reasons)
Lost to follow up (n = …) (give reasons)
Discontinued intervention (n = …) (give reasons)
Analyzed (n = …)
Excluded from analysis (give reasons) (n = …)
Analyzed (n = …)
Excluded from analysis (give reasons) (n = …)
Environment
Allocation
Follow up
Analysis

Fig. 9.1 Revised template of the CONSORT diagram showing the flow of participants through each stage of a randomized trial (Source: http://www.consort-statement.org)

or long follow-up time, or both in order to yield a sufficient number of cases to judge the efficacy of the intervention. One possible strategy to overcome this problem is to use a surrogate endpoint – a short term marker of the disease process – as a substitute for a hard endpoint such as disease incidence or death. For example, a prophylactic vaccine against human papilloma virus given during adolescence is intended to prevent occurrence of cervical cancer in middle age, but its efficacy was evaluated using precancerous endpoints (cervical intraepithelial neoplasia grade 2, grade 3, and adenocarcinoma in situ) (Future II Study Group 2007). It should be noted that cervical cancer is preventable, in the absence of a vaccine, by screening. Hence, it would not be ethical to use cervical cancer as an endpoint.

The promise of surrogate endpoints is that they may make intervention trials smaller, faster, or cheaper. Despite the attractiveness of this promise, the use of surrogate endpoints is fraught with difficulty. Schatzkin and colleagues have,

Table 9.2 Phases of intervention trials

Phase	Aims	Comment
I	Route of administration and dosage: maximally tolerated dose by using a dose escalation scheme	Often on volunteers
II	Evidence of "activity" of our intervention by means of "promising" outcome measures	Better randomized than not
III	Efficacy of an intervention by means of randomized comparisons and a "definite" endpoint	
IV	Effectiveness of proven interventions in wide-scale use, sometimes through post-marketing surveillance	Better randomized than not

in a series of articles, reviewed the problems of using surrogate endpoints in cancer research (Schatzkin et al. 1990, 1996; Schatzkin and Gail 2002). They suggest that there are currently only two clear candidates for surrogate endpoints in cancer research: prevention of HPV infection for subsequent cervical cancer and prevention of colorectal adenomas for subsequent colorectal cancer. Apart from these two examples, they suggest that surrogate endpoints may have more use in phase II trials than phase III trials (for a definition of phase II and phase III trials, see Table 9.2).

The problem of validation of surrogate endpoints is an area of active statistical research. In a seminal paper, Prentice (1989) suggested the following definition of a valid surrogate: it must yield a valid test of the null hypothesis of no association between treatment and the true response. In operational terms, this means that the incidence rate of the true disease must be independent of the treatment history given the current value of the surrogate endpoint. Prentice's criterion is too restrictive to be satisfied in practice as it requires the effect of the treatment on the disease outcome to be mediated entirely through the surrogate endpoints. Some attempts have been made to broaden the definition of a surrogate endpoint by quantifying the proportion of treatment effect explained by a surrogate (Freedman et al. 1992), but this extension has not been widely accepted. A summary of current research is given by the report of an NIH (US National Institutes of Health) workshop on surrogate endpoints (De Gruttola et al. 2001).

9.2.2 From Observation to Intervention

Whereas the choice of intervention in a prevention trial is often suggested by observational studies, the intervention may be a radical simplification of those observations. This is especially true in nutritional epidemiology (see chapter ▶Nutritional Epidemiology of this handbook) in which chemoprevention (e.g., administration of specific vitamins) has often been used as a substitute for dietary modification. This simplification involves an extra level of extrapolation – above issues of timing, dose, and duration – which makes the results of such studies particularly hard to interpret when they contradict observational studies.

Intervention studies can sometimes produce results suggesting that a treatment is harmful, increasing the risk of disease instead of decreasing it. The most notorious example is the use of beta-carotene supplements to prevent lung cancer (see Example 9.1).

Example 9.1. Beta-Carotene and lung cancer: An unexpected harmful effect of treatment

One of the most consistent findings in nutritional epidemiology is the protective effect of fresh fruit and vegetable consumption on cancer risk (World Cancer Research Fund – American Institute for Cancer Research 1997). Peto et al. (1981) put forward the hypothesis that beta-carotene was the active agent responsible for this protective effect, and subsequently a number of cancer chemoprevention trials were conducted using supplementation with beta-carotene as an intervention. The ATBC (The Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study Group 1994) and CARET (Omenn et al. 1996b) trials were two large chemoprevention trials using beta-carotene in men at high risk of lung cancer, i.e., long-term heavy smokers (ATBC/CARET) or asbestos-exposed workers (CARET). The ATBC trial showed a higher incidence of lung cancer among participants receiving beta-carotene compared with those who did not with a relative risk of 1.6 (95% confidence interval CI: 1.02–1.33). Subsequently, the intervention with beta-carotene in the CARET study was stopped 21 months early, after a median of 3.7 years of follow-up because of clear evidence of no benefit and substantial evidence of possible harm. The actively treated group had a relative risk of lung cancer of 1.28 (95% CI: 1.04–1.57) compared with the group receiving placebo. The results of the Physicians' Health Study (PHS) were published at the same time (Hennekens et al. 1996) and showed no effect of beta-carotene on lung cancer risk, with 82 cases of lung cancer in the beta-carotene group and 88 in the placebo group. However, the PHS had low power to detect small changes in lung cancer risk due to the small number of cases.

The trials of beta-carotene relied on a number of extrapolations from observational data. The first extrapolation was that beta-carotene was the active agent in the protective effect observed for fresh fruits and vegetables. Although direct associations with plasma levels of beta-carotene have also been observed, it may be that beta-carotene is acting as a marker for fresh fruit and vegetable consumption and is not the active agent. The second extrapolation concerns the dose level. The dose of beta-carotene given and the median serum beta-carotene concentration achieved in these studies exceeded by many times the level that could be achieved by normal dietary intake (IARC 1998), and it is possible that beta-carotene becomes harmful at such high doses, while remaining protective at doses associated with a healthy diet. Last, but not least, it was assumed that high-risk, middle-aged, individuals, who probably harbored premalignant lesions in the lung, would benefit from the same active substances that were believed to be beneficial in the prevention of early stages of carcinogenesis.

When planning an intervention study, it may be a useful exercise to consider the possible interpretations of adverse effects of treatment. This thought experiment may reveal weaknesses in the motivation for, or design of, the intervention study.

9.3 The Origin of Randomized Trials

The technique of randomization was originally developed in agricultural experiments in the 1920s when individual plots of land were randomized. Trials in humans strictly controlled by random allocation date back to the mid-1940s. Many of the earliest trials were carried out in Britain by the Medical Research Council (Hill 1962). The first trial in which treatment was randomly allocated to individuals, although not the first to be reported (Doll 1998), was a prevention trial designed to test the efficacy of immunization against whooping cough (Medical Research Council Whooping-Cough Immunization Committee 1951). Parents of children aged 6–18 months were asked to volunteer to have their children entered into the trial. They were given a pamphlet describing the study, which included the information that half the inoculations would not be against whooping cough but would be "anti-catarrhal." No child was entered until a parental consent form had been signed.

The spread of randomization, until it became an essential requirement for the licensing of new drugs, was initially slow and not without opposition. Many clinicians considered randomization to be less worthy than the use of criteria to distinguish between individuals who will and will not respond to the intervention. There is currently no competing methodology for randomized controlled trials, although there are attempts to adapt the design and analysis to special contexts and difficulties such as outcomes that are extended in time and trials with informative dropout (Lavori and Kelsey 2002).

9.4 Planning of Trials

Trials are traditionally classified into four phases (Table 9.2). In this chapter, we will be dealing only with issues related to phase III and phase IV randomized trials. For design issues of phase I and phase II trials, readers are referred elsewhere (Simon 2001).

The organization of a trial requires careful advance planning. This is particularly true for multicentric trials, which have become increasingly common. The aims and methods of the trial should be described in detail in a protocol document. This will contain the scientific background of the problem under study. In topics where a substantial amount of work has already been done, a systematic review (meta-analysis, see chapter ▸Meta-Analysis in Epidemiology in this handbook) of the outcomes of published randomized trials on the same type of intervention is highly preferable to a narrative review. A meta-analysis will, in fact, greatly help to evaluate the consistency of the methods and results of previous work, any avoidable pitfalls in study design, and the most likely effectiveness of the intervention under study. The latter information allows the sample size for the trial to be calculated (see chapter ▸Sample Size Determination in Epidemiological Studies of this handbook). The protocol should also include clear statements about (1) preventive measures to be used (both intervention and control); (2) types of individuals or groups to be admitted (participants); (3) assessment of response (endpoints); (4) entry criteria and treatment allocation; (5) exclusions, withdrawals, and protocol departures; (6) size and duration of the trial; (7) strategy for the statistical analysis; and (8) ethical aspects.

When the protocol has been written, information about the trial design should be deposited into a clinical trials registry before the onset of patient enrollment. The World Health Organization sets standards for clinical trial registries through the International Clinical Trials Registry Platform (http://www.who.int/ictrp). It stipulates a list of 20 items that must be given for a trial to be considered fully registered.

The International Committee of Medical Journal Editors has stipulated a policy requiring preregistration as a requirement for publication of a trial's findings (DeAngelis et al. 2004). This policy is designed to prevent selective reporting of trial outcomes. However, a recent systematic review shows that preregistration of trials still poorly adhered to Mathieu et al. (2009).

9.5 Definition of the Intervention

The effectiveness of interventions to be compared in randomized primary prevention trials are usually known in broad terms from the outset. Many questions of primary prevention have first been evaluated in trials of treatments or secondary prevention. For example, the Physicians' Health Study (Steering Committee of the Physicians' Health Study Research Group 1989) demonstrated that a daily low-dose (325 mg) aspirin reduced the risk of first myocardial infarction by 44%. But when the trial was begun, there was already substantial statistical evidence for the preventive efficacy of aspirin from secondary prevention trials of those who had already experienced a cardiovascular event (ISIS-2 (Second International Study of Infarct Survival) Collaborative Group 1988). Similarly, the suggestion to use tamoxifen or the new selective estrogen receptor modulators (e.g., raloxifene) in the primary prevention of breast cancer came from the observation that tamoxifen reduced the incidence of contralateral breast cancer when used as an adjuvant therapy in women with breast cancer (Early Breast Cancer Trialists' Collaborative Group 1998). Such previous experiences in "diseased" people have generally established a "safety profile" and enabled exclusion or estimation of side effects that occur once in thousands (if not necessarily in ten thousands) of recipients of a certain type of treatment. Other times, when the intervention consists in behavioral changes, prior evidence for the benefits of certain lifestyle modifications (e.g., smoking cessation, dietary changes, adoption of safe sex behavior) has come from a consistent body of knowledge on risk factors accumulated in a large variety of contexts (e.g., ecological studies, studies of observational epidemiology or, as it is the case for some infectious agents, clear knowledge on the routes of transmission of the infections).

Based on this background knowledge, two major questions need to be addressed in order to identify unambiguously the type of intervention study. The first question concerns what is being compared with what. The basic design for primary prevention trials is a two-arm comparison. As discussed by Green (2002), however, this simple structure can encompass a variety of different comparisons, as shown in Table 9.3.

Table 9.3 Possible types of comparison in intervention trials

Intervention vs	No intervention
	Placebo
	Another intervention
	Same intervention at a different dose (or duration)
	Same intervention, but later (only for participants who experience a certain event)

The feasibility and necessity to include a placebo arm depend on the nature of the intervention and of the outcome measure. When two drugs are being compared, it is generally easy to create a placebo, provided that the route of administration is an oral one. In therapeutic trials, the use of a non-active compound has caused ethical concerns. A comparison of new treatment with old treatment is often more appropriate. As noted by Temple and Ellenberg (2000), however, placebo controls are ethical if there are no permanent adverse consequence of delaying or omitting available treatment and if the participants are fully informed about their alternatives. These favorable conditions typically apply to preventive trials. For intra-muscular treatments (e.g., vaccines), the use of one or more "dummy injections" is hard to justify. It has, however, been advocated in special circumstances. In a randomized trial of a prophylactic vaccine against HPV type 16, a sexually transmitted virus which is now considered a necessary cause of cervical cancer, three injections of the aluminum adjuvant used in the active vaccine were deemed justifiable in order to avoid possible imbalances in the sexual behavior of the young female participants and to evaluate side effects attributable to the active agent (i.e., virus-like particles) (Koutsky et al. 2002). Indeed, the percentage of women who discontinued the study owing to an adverse effect of treatment (0.4%) was the same in the vaccine arm and the placebo arm (Koutsky et al. 2002).

One way to avoid the use of placebo is to compare different doses or durations of a specific intervention. This is typically feasible when some type of drug or dietary supplement, whose minimal effective dose is unclear, is under evaluation. More frequently, however, in prevention trials, it is the intensity of the intervention that can be modulated. In behavioral intervention trials, it is often very useful to compare a labor-intensive and expensive package for smoking cessation or dietary changes with some simple and inexpensive message at a community or individual level (e.g., pamphlets, simple recommendations from a general practitioner, etc.).

Finally, the timing of the intervention can be randomized into an early versus delayed intervention trial. This approach is very often indicated in therapeutic trials (e.g., chemo- or radiotherapy at primary cancer diagnosis vs the same therapy at cancer recurrence), but it has some appeal in certain prevention trials as well. An example is provided by cervical cancer screening, which is currently based on use of the Papanicolaou (Pap) smear to detect abnormal cells that may indicate precancerous lesions on the cervix. The majority of such lesions regress without intervention, but a few may progress to cervical cancer over the course of several years. HPV testing is now considered a more sensitive test than Pap smear in the detection of precancerous lesions (Cuzick et al. 2000). Its use is already approved in the triage of cytological abnormalities. Thus, an appropriate design for testing the efficacy of HPV testing in cervical cancer screening would be to use Pap smear in both arms and compare the concurrent use of HPV testing and Pap smear in one arm with delayed HPV testing in the other arm among women with abnormal Pap smear findings.

9.6 Factorial Design

The factorial design represents an efficient alternative to two-arm comparisons. The simplest design is the balanced 2×2 factorial but factorial designs can be generalized to more than two dimensions (Table 9.4). In a factorial design, two or more interventions are simultaneously tested in the same population. Allocation of interventions is carried out in such a way that there is no association between different interventions in the study population, and therefore no confounding, under the assumption that there is no interaction between interventions. As noted by Armitage and Berry (1987), this design contravenes a good principle of experimentation, namely, that only one factor should be changed at a time. The principal advantage of the factorial design is its ability to answer two or more questions in a single trial.

Table 9.4 Illustration of the factorial design with 8,000 participants and three possible treatments. As the number of treatments simultaneously under test increases, the number of participants receiving each combination of treatment diminishes, but the number receiving any given treatment (e.g., treatment A) is always 4,000

	Number of	Treatment		
Arm	participants	A	B	C
Two arm trial of treatment A				
1	4,000	Yes		
2	4,000	No		
2 × 2 trial of treatments A and B				
1	2,000	Yes	Yes	
2	2,000	Yes	No	
3	2,000	No	Yes	
4	2,000	No	No	
2 × 2 × 2 trial of treatments A, B and C				
1	1,000	Yes	Yes	Yes
2	1,000	Yes	Yes	No
3	1,000	Yes	No	Yes
4	1,000	Yes	No	No
5	1,000	No	Yes	Yes
6	1,000	No	Yes	No
7	1,000	No	No	Yes
8	1,000	No	No	No

In a recent trial of 20,536 UK adults with occlusive arterial disease or diabetes (Heart Protection Study Collaborative Group 2002a, b), participants were randomly allocated in a factorial design to receive 40 mg of simvastatin (a cholesterol-lowering agent) or placebo, and antioxidant vitamin supplementation or placebo. These two interventions seemed to be, at the time the study was started, equally promising. It would have been possible to evaluate the two interventions separately in two independent trials, but this would have required twice the sample size, or over 40,000 participants. A highly significant 18% proportional reduction in the coronary death rate was found among participants who had received simvastatin. This beneficial effect was reflected also in an overall mortality reduction compared to the placebo group. Conversely, antioxidant vitamins did not produce any significant reduction in the 5-year mortality from, or incidence of, any type of vascular disease, cancer, or other major outcome. An asset of the factorial design is also the opportunity to evaluate interactions, i.e., whether two interventions in combination differ, with respect to efficacy or side effects, from either intervention alone. In the HPSCG trial, for instance, the efficacy of simvastatin was not modified by antioxidant vitamins.

The factorial design, on account of the corresponding gain in cost efficiency, is especially valuable in prevention trials that are, on average, much larger than therapeutic trials. A factorial design is often excluded out of fear of complicating trial operations. In fact, the randomization process can be easily adapted to allocate participants to different combinations of interventions. As mentioned by Buring (2002), it is, however, essential that none of the interventions under evaluation (1) complicates eligibility criteria by, for instance, important contraindications for certain interventions or (2) causes any side effect that could lead to poor compliance or loss to follow-up.

9.7 Definition of Participants

Eligibility criteria in intervention studies must aim at three things: (1) optimizing the potential benefit to the participants while minimizing the risk of adverse effects, (2) enrolling participants who are likely to adhere to the intervention and follow-up requirements, and most importantly (3) including a sufficiently large number of participants to produce unambiguous results even if the benefit of the intervention is small. In many types of trials, but especially in preventive trials, broad eligibility criteria are desirable because they can simplify enrolling procedures and avoid the need for complicated and expensive tests at study entry (see Box 9.1). They also increase the generalizability of the results.

A “run-in” period can be implemented to allow potential participants who have difficulties adhering to the protocol to withdraw prior to randomization. As discussed by Buring (2002), the format of the run-in period depends on the nature of the trial. In trials of pill-taking regimens, as in the HPSCG (Heart Protection Study Collaborative Group 2002a, b) factorial trial of simvastatin and antioxidant vitamins, the run-in period involved a few-week period of placebo to allow review of blood exams, followed by a few weeks of active treatment to allow a pre-randomization assessment of LDL (low-density lipoprotein)-lowering responsiveness of each individual and to exclude major adverse effects. In studies of behavioral interventions, the run-in period may consist of attendance at visits

Box 9.1. Large, simple randomized trials

Many lives could be saved by moderate reductions in the common causes of death.

In a series of articles, Peto and colleagues (Collins et al. 1987; Peto et al. 1995; Peto and Baigent 1998; Yusuf et al. 1984) have promoted the idea of large, simple randomized trials to investigate the benefits of widely practicable treatments for common conditions. The essence of their argument is that only large, randomized trials can answer questions about moderate health benefits in a way that is free of bias and the play of chance. Moderate reductions in mortality may correspond to a large number of deaths prevented if the condition is common and the treatment is widely available. For example, 100,000 deaths per year could be prevented or substantially delayed in developed countries by routine use of antiplatelet therapy in all patients with clinical evidence of occlusive vascular disease. This reduction corresponds to a 10% reduction in all vascular deaths in the age range 35–69 (Antiplatelet Trialists' Collaboration 1994).

Examples of the large randomized trials promoted by Peto and colleagues include the ISIS (International Study of Infarct Survival) trials (ISIS-1 (First International Study of Infarct Survival) Collaborative Group 1986; ISIS-2 (Second International Study of Infarct Survival) Collaborative Group 1988; ISIS-3 (Third International Study of Infarct Survival) Collaborative Group 1992; ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group 1995) in which tens of thousands of participants were randomized. The need to randomize such large numbers of subjects imposes some design constraints on the trial. In particular, the entry criteria, treatment, and data requirements must all be greatly simplified. In order to simplify the entry criteria into large trials, Peto et al. (1995) have proposed an "uncertainty principle," which states that the sole eligibility criterion for entry is that both patient and doctor should be substantially uncertain about the appropriateness for this particular patient of each of the trial treatments (a more complete statement is given by Peto and Baigent (1998)). Broad eligibility criteria can simplify enrolling procedures and avoid the need for complicated and expensive tests at study entry.

The principle of conducting large, simple randomized trials is not appropriate for the development of novel drugs. In this context, an extensive regulatory framework has developed, which is summarized by the guidelines of the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (http://www.ich.org).

or laboratory procedures, including completion of forms similar to those that would be used in the actual trial. Restricting a prevention trial to proven good compliers may result in a pool of participants that differs from the general population with respect to outcomes. This problem may be perceived as a loss of external validity or generalizability of trial findings. As noted, however, by Hennekens and Buring (1998), the primary requirement of a generalizable study is internal validity, which may in fact be increased by the exclusion of poor compliers.

A more practical way to improve the generalizability of the results of a trial is to broaden eligibility criteria as much as possible, thus allowing the benefits of trial participation to be available more widely across populations. In the HPSCG study, for instance, substantial benefit was demonstrated not only in those already known to have had coronary disease, but also in those without diagnosed coronary disease who had cerebrovascular disease, peripheral arterial disease, or diabetes, irrespective of the blood lipid concentrations when treatment was initiated. Widespread implementation of these findings on the basis of some clinical diagnoses would therefore be relatively straightforward, without the need for extensive screening in the general population. Finally, broad eligibility criteria can also allow to enroll a larger number of participants and, in some instances, some cautious subgroup analysis.

An eligibility criterion which, however, allows substantial efficiency gain, and which must therefore be seriously considered, is high risk for the disease meant to be prevented. Age, sex, or family or personal history might be considered to identify such individuals. Since the power of the study is proportional to the number of endpoints, not simply to the number of participants enrolled, an intervention study which includes high-risk individuals will be small and of shorter duration (and, as a consequence, cheaper and likelier to be accomplished) than a study which includes lower risk participants. The demonstration of a benefit among those with high-risk characteristics will have to be applied, however, with great caution to a more heterogeneous population.

9.8 Enrollment

Enrollment of thousands participants in prevention trials is a major challenge and deserves to be monitored carefully. Since prevention trials do not require diseased people, they rarely use hospitals as sources of participants. Outpatient clinics are, however, used in some secondary prevention studies.

A preliminary question in community-based enrollment is whether trial participants can be volunteers, provided they meet the eligibility criteria and can be recruited in the required number or whether it is necessary to extend an invitation to join the study to all eligible persons in a predefined area. If the former applies, local media, pamphlet distribution, and direct contacts with special associations can be used in order to invite as many participants as required. In the previously mentioned trial of a vaccine against HPV 16, for instance, young women were recruited in the United States through advertisements on college campuses and in the surrounding communities (Koutsky et al. 2002). In order to diminish the probability of enrolling women who were infected with HPV16, only women who reported that they had no more than five male sex partners during their lifetime were eligible for participation. Obviously, when participants are volunteers recruited from a population or subpopulation that has not been enumerated, it will be impossible to evaluate accurately the fraction of the persons who eventually opted for participating in the trial, let alone any difference between participants and non-participants. This approach is, however, the only possibility when participants are defined according to lifestyle characteristics (e.g., current smoking habit, overweight, sleeping disorders, sexual habits, etc.) that cannot be derived from administrative records.

In other instances, however, it may be important to have access to the entire list of the population of a predefined area and to make an effort to achieve, as much as possible, high participation or, at least, to estimate the participation fraction. This is necessary when the results of the trial must be generalizable to the whole of the host population and in particular when the trial must include certain subgroups. In developed countries, complete and fairly well updated population lists exist, such as censuses, electoral rolls, and general practitioners' lists. In the latter case, a direct involvement of general practitioners in the recruitment of their own patients is known to improve the yield of participants (Knatterud 2002).

In most developing countries, especially in the many areas where substantial migrations from rural to urban areas are ongoing, reliable population lists are seldom available, and a door-to-door enumeration of the target population of the trials is generally a prerequisite of the recruitment exercise. The complete enumeration of the target population is especially important in the evaluation of interventions such as immunization programs or screening programs, where a good coverage is essential for the efficacy of the intervention. The participation fraction (i.e., the acceptance of the intervention) becomes part of the evaluation of the intervention under study.

9.9 Randomization

An eligible participant should be admitted formally to the trial, by entry of his or her name into a register and the allocation of a unique identification number. The random allocation of the intervention should be determined after entry or a run-in period. The fundamental reason for random allocation is to maximize the likelihood that participants with similar characteristics will be allocated in similar proportions to the different interventions being investigated. The decision to enter a participant must be made irreversibly regardless of which trial arm the participant will be allocated. This precludes systematic allocation systems and discourages easily guessable coding systems, such as one in which an active intervention is coded as A and the placebo as B, which may lead to unblinding of the study. The bias inherent in non-randomized studies may be more severe in therapeutic trials (where diseased patients can differ substantially according to prognostic factors, at least in the clinician's opinion) but is not negligible in preventive trials as well.

A common allocation technique, discussed by Knatterud (2002), is to generate a separate randomization schedule for each of the trial sites and, within each site, to have some blocking so that, after a fixed number of participants enrolled, there will be equal numbers assigned to each trial arm. It is advisable to make smaller blocks at the beginning, when the enrollment capacity of trial sites is unknown. Ideally, there should be a central office where the randomization is carried out and which enforces good record keeping and adherence to the protocol. A telephone call to the randomization office can include some verification of the patients eligibility before randomization. With current technology, it is possible to implement an automated menu-driven telephone system that is available 24 h a day, 7 days a week (Heart Protection Study Collaborative Group 2002b). Elaborate "stratification" schemes in which a separate randomization block is made for participants with different characteristics are not encouraged in large trials (Peto et al. 1976). Proper statistical methods (see chapter ▸Regression Methods for Epidemiological Analysis of this handbook) can make due allowance, when comparing interventions, for what was initially known about each participant.

9.10 Cluster Randomization

Interventions in communities or other groups have frequently been investigated without using randomization. Longitudinal studies of the health consequences of water fluoridation (Horowitz 1996) and initial cardiovascular disease prevention studies such as the North Karelia Project (Puska et al. 1976) provided early evidence that community-level health interventions could benefit large groups of individuals. However, non-randomized community intervention trials may be confounded by secular trends. For example, in the two studies cited above, the incidence of dental caries and cardiovascular disease may have diminished due to widespread use of fluoridated toothpaste and improvements in the treatment of hypertension, respectively.

The use of randomization is, therefore, just as important for community trials as for individual-level studies, but it is not appropriate to randomize individuals when the intervention is defined at the group level. Instead, groups of individuals or "clusters" are randomized. A cluster may be defined geographically (e.g., cities, countries, villages) or otherwise (e.g., workplaces, schools, clinical practices) (Atienza and King 2002). Cluster randomization can also avoid the potential for contamination between interventions, which may occur in individually randomized trials if the intervention involves education or behavior modification. Furthermore, cluster randomization by village improves the chances of long duration follow-up in developing countries where personal identifiers, such as name and date of birth, are unreliable (Gambia Hepatitis Study Group 1987), but strong community links exist.

The main consequence of randomizing clusters is that the outcomes for trial participants in the same cluster are no longer independent. This lack of independence has implications for both analysis and design. The analysis of the study must take into account the presence of the clusters (for a detailed discussion of cluster randomized trials, see also chapter ▸Cluster Randomized Trials of this handbook).

Various techniques for doing so are reviewed by Donner (1998). If the clustering effect is ignored, as in 70% of cluster trials reviewed by Divine et al. (1992), p-values will be too small and exaggerate the statistical significance of the results. Likewise, confidence intervals will be too narrow. Even if a correct statistical analysis is used, the power of a cluster randomized trial is reduced compared to an individually randomized trial. When the outcome of the trial is the mean of a continuous variable (e.g., serum cholesterol level), the increase in sample size needed to maintain the same power as an individually randomized study is given by the formula $1 + (m - 1)r$, where m is the number of participants per cluster and r is the intracluster correlation coefficient (Kerry and Bland 1998), which measures the proportion of variation of the outcome in the study population that is due to differences between clusters. Typical values of r are very close to zero, but when the number of participants per cluster (m) is large, there may be a serious loss of efficiency compared with an individually randomized study.

A large number of clusters also increases the possibility that the randomization will produce balanced intervention groups compared to trials where a small number of clusters has been randomly allocated. "Restricted randomization" after matching or stratifying communities according to selected factors has been attempted. In the Community Intervention Trial for Smoking Cessation (COMMIT Research Group 1991), for instance, investigators examined 11 pairs of communities that were matched on sociodemographic factors. Matching is only justified for variables that are strongly related to the outcome of interest (Klar and Donner 1997). The same principle applies to sampling of controls in observational studies (see chapter ▸Case-Control Studies of this handbook).

Notwithstanding unsolved issues, cluster design can greatly contribute to the evaluation of preventive strategies and methodological advances should be pursued.

9.11 Outcomes

In the design of prevention trials, a distinction is generally made between primary and secondary outcomes (Anderson and Prentice 1999). The primary outcome is typically the clinical disease to be prevented or controlled that provides the central justification for the trial and determines the study size. Secondary outcomes are disease events that also motivated the trial, but that by themselves would be unlikely to justify a full-scale intervention.

The outcome measurements could be influenced by a knowledge of which intervention was used, thus producing serious bias. In a single-blinded trial, the identity of the allocated treatment is concealed from the participants. A double-blinded trial is one in which the doctor, or other technical expert, who assesses response also is unaware of the intervention identities. As discussed by Green (2002), it may not be desirable to extend blinding to the independent data and safety monitoring board (DSMB). The DSMB, as often stated in the participant's consent form, is the watcher of accumulating data and needs, therefore, to be aware of the allocation.

Unlike therapeutic trials, large prevention trials cannot rely on outcomes that require frequent clinical visits for specialized tests or potential adverse reactions. Indeed, the safety of any intervention has to be documented before the planning of any large preventive trial. Furthermore, even if participants in preventive trials are selected on the basis of elevated risk for the diseases that are targeted for the prevention, primary outcome events may constitute a small minority of the disease events experienced by study participants during the course of the trial and perhaps even a small minority of disease events that may in some way be affected by intervention activities. Hence, there is an obligation to define sets of outcomes to be fully ascertained, including often overall mortality, that provide an opportunity to assess the overall risks and benefits in the target population.

The International Breast Cancer Intervention Study (IBIS-I), for instance, was a double-blind placebo-controlled randomized trial of tamoxifen, 20 mg/day for 5 years, in 7,152 women at increased risk of breast cancer (IBIS Investigators 2002). The primary outcome measure was the frequency of breast cancer, but outcomes other than breast cancer were found to be very important, too. A 32% reduction (95% CI: 8–50%) in the primary outcome was found. It was accompanied, however, by a 2.5-fold increase (95% CI: 1.5–4.4) in thromboembolic events. Eventually, a significant excess of deaths from all causes was observed in the tamoxifen group (25 vs 11, $p = 0.028$). The conclusions of IBIS-I were that the overall risk to benefit ratio for the use of tamoxifen in prevention was still unclear and continued follow-up of the trial was essential.

Although sample size determination is dealt with elsewhere in this volume (chapter ▸Sample Size Determination in Epidemiological Studies), the dangers of inadequate statistical power on intervention studies cannot be overemphasized. More than therapeutic trials, preventive trials cannot be easily replicated, for economical and logistic reasons, and collaborative reanalyses can seldom remedy at the lack of a definitive answer from a specific trial. As discussed by Buring (2002), every effort should be made during the planning phase of the trial to choose an adequate sample size and length of follow-up. Secular declines in disease rates within the general population and the failure to achieve a sufficient sample size or to accrue sufficient endpoints are frequent in prevention trials. Extending the duration of the trial is often a good option to achieve a more definitive result for a relatively small increase in total cost.

9.12 Follow-Up, Exclusions, and Withdrawal

Various aspects of the statistical analysis of follow-up data from intervention trials are dealt with elsewhere in this volume (chapters ▸Analysis of Continuous Covariates and Dose-Effect Analysis, ▸Regression Methods for Epidemiological Analysis, ▸Survival Analysis and ▸Missing Data).

The series of articles by Peto et al. (1976, 1977) gives a clear and comprehensive overview of the analysis of long-term follow-up in randomized controlled trials. Survival analysis is the standard method of analyzing time-to-event data because it makes use of the full information (i.e., not just which events occurred, but also when they occurred) and can take into account right censoring, which occurs when the study ends or when participants are lost to follow-up. Interval censoring can also occur when the events under study are only known to have occurred in some interval in time, generally between two follow-up visits. For example, if the event is not death but some measurement that requires a clinical examination (e.g., the appearance or the disappearance of HPV infection, which is asymptomatic and requires a specimen of exfoliated cervical cells for detection), the exact time of the event is unknown, except that it occurred between two visits. Appropriate statistical methods can take this uncertainty into account.

Censoring resulting from losses and dropouts is problematic if the censoring is "informative" (i.e., related to outcome), a situation that is difficult to rule out. As noted by Peto et al. (1977), "rigorous entry criteria are not necessary for a randomized trial, but rigorous follow-up is. Even patients who do not get the proper intervention must not be withdrawn from the analysis." Often referred to as analysis by intention to treat, this approach is the only one that provides a valid answer to a real question. It tests the "policy" (or intention) to be evaluated at the time of randomization. Every possible effort must therefore be made, at the level of trial design and implementation, to identify and, as much as possible, avoid any causes of non-adherence. This may entail a run-in period and a simplified follow-up protocol.

A special problem, different from the one of losses at follow-up, is represented by participants who may be discovered to contravene the eligibility criteria after randomization. In a double-blind trial of a prophylactic vaccine against HPV 16, 2,392 young women were randomized (Koutsky et al. 2002). Since the tested vaccine was not supposed to work among women who were already infected with HPV 16, 36% of the randomized trial participants were subsequently excluded, mainly because HPV tests revealed that they were already infected at enrollment. In principle, no bias should have been introduced since the results of HPV tests became available only after the randomization. An alternative would have been to postpone the randomization, but this was considered impractical. The trial by Koutsky et al. (2002), however, was the first test of a vaccine against HPV. It was not meant to test a "policy" of vaccination in the general population, but the efficacy of a new vaccine under optimal conditions (i.e., among women unexposed to HPV). A larger, population-based trial of HPV vaccine would be necessary to evaluate the effectiveness of HPV vaccine as a tool for cervical cancer control (Plummer and Franceschi 2002). Analysis of such a trial by intention to treat would give the programmatic efficacy of the vaccine program, which is distinct from the efficacy of the vaccine itself.

Finally, a special challenge is represented by the follow-up of community-based health interventions. As discussed by Atienza and King (2002) and Koepsell et al. (1992), it is typically not possible to assess all individuals of interest in the selected communities. Two main approaches to obtaining these longitudinal individual-level data are (1) to follow-up groups of individuals over time and (2) to assess different cross-sections in each time period. Several large-scale community-based interventions (e.g., COMMIT) have utilized both approaches, cognizant of different strengths and weaknesses (Koepsell et al. 1992).

9.13 Conclusions

The randomized controlled trial is one of the most powerful tools available in epidemiology. When properly conducted and reported, it allows the evaluation of a disease risk factor that is free from bias and may also, with an appropriate design, give a quantitative estimate of the public health benefit that may be expected from an intervention.

Even well-conducted randomized controlled trials have problems of interpretation. These problems center on the generalizability of the findings and concern not only the selection of the study participants, but also the timing, dose, and duration of the intervention and the length of time over which a beneficial effect was observed. If there are no such problems, the results of randomized controlled trials may be taken to provide a definitive answer and, in particular, overrule the results from observational studies when these disagree. In general, however, the results of intervention trials must be considered as part of the spectrum of available evidence that includes observational studies in humans and experimental data that provide mechanistic evidence.

References

Alpha-Tocopherol Beta Carotene Cancer Prevention Study Group (1994) The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. N Engl J Med 330:1029–1035

Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gotzsche PC, Lang T (2001) The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 134:663–694

Anderson GL, Prentice RL (1999) Individually randomized intervention trials for disease prevention and control. Stat Methods Med Res 8:287–309

Antiplatelet Trialists' Collaboration (1994) Collaborative overview of randomised trials of antiplatelet therapy–I: prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients. BMJ 308:81–106

Armitage P, Berry G (1987) Statistical methods in medical research. Blackwell, Oxford

ATBC (Alpha-Tocopherol Beta Carotene) Cancer Prevention Study Group (1994) The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. Ann Epidemiol 4:1–10

Atienza AA, King AC (2002) Community-based health intervention trials: an overview of methodological issues. Epidemiol Rev 24:72–79

Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF (1996) Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 276:637–639

Bonithon-Kopp C, Kronborg O, Giacosa A, Rath U, Faivre J, for the European Cancer Prevention Organisation Study Group (2000) Calcium and fibre supplementation in prevention of colorectal adenoma recurrence: a randomised intervention trial. Lancet 356:1300–1306

Buring JE (2002) Special issues related to randomized trials of primary prevention. Epidemiol Rev 24:67–71

CAST (Chinese Acute Stroke Trial) Collaborative Group (1997) CAST: randomised placebo-controlled trial of early aspirin use in 20000 patients with acute ischaemic stroke. Lancet 349:1641–1649

Collins R, Gray R, Godwin J, Peto R (1987) Avoidance of large biases and large random errors in the assessment of moderate treatment effects: the need for systematic overviews. Stat Med 6:245–254

COMMIT Research Group (1991) Community Intervention Trial for Smoking Cessation (COMMIT): summary of design and intervention. J Natl Cancer Inst 83:1620–1628

Cuzick J, Sasieni P, Davies P, Adams J, Normand C, Frater A, van Ballegooijen M, van den Akker-van Marle E (2000) A systematic review of the role of human papilloma virus (HPV) testing within a cervical screening programme: summary and conclusions. Br J Cancer 83:561–565

De Gruttola VG, Clax P, DeMets DL, Downing GJ, Ellenberg SS, Friedman L, Gail MH, Prentice R, Wittes J, Zeger SL (2001) Considerations in the evaluation of surrogate endpoints in clinical trials. Summary of a National Institutes of Health workshop. Control Clin Trials 22:485–502

DeAngelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, Kotzin S, Laine C, Marusic A, Overbeke AJ, Schroeder TV, Sox HC, Van Der Weyden MB, International Committee of Medical Journal Editors. (2004) Clinical trial registration: a statement from the International Committee of Medical Journal Editors. JAMA 292:1363–1364

Divine GW, Brown JT, Frazier LM (1992) The unit of analysis error in studies about physicians' patient care behavior. J Gen Intern Med 7:623–629

Doll R (1998) Controlled trials: the 1948 watershed. BMJ 317:1217–1220

Doll R (2002) Proof of causality: deduction from epidemiological observation. Perspect Biol Med 45:499–515

Doll R, Peto R (1981) The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. J Natl Cancer Inst 66:1191–1308

Donner A (1998) Some aspects of the design and analysis of cluster randomized trials. Appl Stat 47:95–113

Early Breast Cancer Trialists' Collaborative Group (1998) Tamoxifen for early breast cancer: an overview of the randomised trials. Lancet 351:1451–1467

Future II Study Group (2007) Effect of prophylactic human papillomavirus L1 virus-like-particle vaccine on risk of cervical intraepitheliala neoplasia grade 2, grade 3, and adenocarcinoma in situ: a combined analysis of four randomized clinical trials. Lancet 369:1861–1868

Freedman LS, Graubard BI, Schatzkin A (1992) Statistical validation of intermediate endpoints for chronic diseases. Stat Med 11:167–178

Gambia Hepatitis Study Group (1987) The Gambia Hepatitis Intervention Study. Cancer Res 47:5782–5787

Green SB (2002) Design of randomized trials. Epidemiol Rev 24:4–11

Heart Protection Study Collaborative Group (2002a) MRC/BHF Heart Protection Study of antioxidant vitamin supplementation in 20536 high-risk individuals: a randomised placebo-controlled trial. Lancet 360:23–33

Heart Protection Study Collaborative Group (2002b) MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20536 high-risk individuals: a randomised placebo-controlled trial. Lancet 360:7–22

Hennekens CH, Buring JE (1998) Validity versus generalizability in clinical trial design and conduct. J Card Fail 4:239–241

Hennekens CH, Buring JE, Manson JE, Stampfer M, Rosner B, Cook NR, Belanger C, LaMotte F, Gaziano JM, Ridker PM, Willett W, Peto R (1996) Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease. N Engl J Med 334:1145–1149

Hill AB (1962) Statistical methods in clinical and preventive medicine. Livinstone, Edinburgh

Horowitz HS (1996) The effectiveness of community water fluoridation in the United States. J Public Health Dent 56:253–258

IARC (1998) Carotenoids. IARC handbooks of cancer prevention 2. International Agency for Research on Cancer, Lyon

IARC (2007) Tobacco control: Reversal of risk after quitting smoking. IARC handbooks of cancer prevention 11. International Agency fo Research on Cancer, Lyon

IBIS Investigators (2002) First results from the International Breast Cancer Intervention Study (IBIS-I): a randomised prevention trial. Lancet 360:817–824

ISIS-1 (First International Study of Infarct Survival) Collaborative Group (1986) Randomised trial of intravenous atenolol among 16 027 cases of suspected acute myocardial infarction: ISIS-1. Lancet 2:57–66

ISIS-2 (Second International Study of Infarct Survival) Collaborative Group (1988) Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17187 cases of suspected acute myocardial infarction: ISIS-2. Lancet 2:349–360

ISIS-3 (Third International Study of Infarct Survival) Collaborative Group (1992) ISIS-3: a randomised comparison of streptokinase vs tissue plasminogen activator vs anistreplase and of aspirin plus heparin vs aspirin alone among 41299 cases of suspected acute myocardial infarction. Lancet 339:753–770

ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group (1995) ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58050 patients with suspected acute myocardial infarction. Lancet 345:669–685

Jacobs ET, Giuliano AR, Roe DJ, Guillen-Rodriguez JM, Hess LM, Alberts DS, Martinez ME (2002) Intake of supplemental and total fiber and risk of colorectal adenoma recurrence in the wheat bran fiber trial. Cancer Epidemiol Biomarkers Prev 11:906–914

Kerry SM, Bland JM (1998) The intracluster correlation coefficient in cluster randomisation. BMJ 316:1455

Klar N, Donner A (1997) The merits of matching in community intervention trials: a cautionary tale. Stat Med 16:1753–1764

Knatterud GL (2002) Management and conduct of randomized controlled trials. Epidemiol Rev 24:12–25

Koepsell TD, Wagner EH, Cheadle AC, Patrick DL, Martin DC, Diehr PH, Perrin EB, Kristal AR, Allan-Andrilla CH, Dey LJ (1992) Selected methodological issues in evaluating community-based health promotion and disease prevention programs. Annu Rev Public Health 13:31–57

Koutsky LA, Ault KA, Wheeler CM, Brown DR, Barr E, Alvarez FB, Chiacchierini LM, Jansen KU (2002) A controlled trial of a human papillomavirus type 16 vaccine. N Engl J Med 347:1645–1651

Lavori P, Kelsey J (eds) (2002) Clinical trials. Epidemiol Rev 24:1–90

Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P (2009) Comparison of registered and published primary outcomes in randomized controlled trials. J Am Med Assoc 302:977–984

Medical Research Council Whooping-Cough Immunization Committee (1951) Prevention of whooping-cough by vaccination. Br Med J 1:1463–1471

Moher D, Schulz KF, Altman DG (2001) The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 357:1191–1194

Omenn GS, Goodman G, Thornquist M, Barnhart S, Balmes J, Cherniack M, Cullen M, Glass A, Keogh J, Liu D, Meyskens F Jr, Perloff M, Valanis B, Williams J Jr (1996a) Chemoprevention of lung cancer: the beta-Carotene and Retinol Efficacy Trial (CARET) in high-risk smokers and asbestos-exposed workers. IARC Sci Publ 67–85

Omenn GS, Goodman GE, Thornquist MD, Balmes J, Cullen MR, Glass A, Keogh JP, Meyskens FL, Valanis B, Williams J H, Barnhart S, Hammar S (1996b) Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. N Engl J Med 334: 1150–1155

Peto R, Baigent C (1998) Trials: the next 50 years. Large scale randomised evidence of moderate benefits. BMJ 317:1170–1171

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG (1976) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer 34:585–612

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples. Br J Cancer 35:1–39

Peto R, Doll R, Buckley JD, Sporn MB (1981) Can dietary beta-carotene materially reduce human cancer rates? Nature 290:201–208

Peto R, Collins R, Gray R (1995) Large-scale randomized evidence: large, simple trials and overviews of trials. J Clin Epidemiol 48:23–40

Plummer M, Franceschi S (2002) Strategies for HPV prevention. Virus Res 89:285–293

Prentice RL (1989) Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med 8:431–440

Puska P, Koskela K, Pakarinen H, Puumalainen P, Soininen V, Tuomilehto J (1976) The North Karelia Project: a programme for community control of cardiovascular diseases. Scand J Soc Med 4:57–60

Schatzkin A, Gail M (2002) The promise and peril of surrogate end points in cancer research. Nat Rev Cancer 2:19–27

Schatzkin A, Freedman LS, Schiffman MH, Dawsey SM (1990) Validation of intermediate end points in cancer research. J Natl Cancer Inst 82:1746–1752

Schatzkin A, Freedman LS, Dorgan J, McShane LM, Schiffman MH, Dawsey SM (1996) Surrogate end points in cancer research: a critique. Cancer Epidemiol Biomark Prev 5:947–953

Simon R (2001) Clinical trials in cancer. In: DeVita VT, Hellman S, Rosenberg SA (eds) Cancer: principles and practice of oncology. LWW, Philadelphia, p 521

Steering Committee of the Physicians' Health Study Research Group (1989) Final report on the aspirin component of the ongoing Physicians' Health Study. N Engl J Med 321:129–135

Stewart BW, McGregor D, Kleihues P (eds) (1996) Principles of chemoprevention, IARC Scientific Publications 139. International Agency for Research on Cancer, Lyon

Temple R, Ellenberg SS (2000) Placebo-controlled trials and active-control trials in the evaluation of new treatments. Ann Intern Med 133:455–463

World Cancer Research Fund – American Institute for Cancer Research (CRF-AICR) (1997) Food, nutrition and the prevention of cancer: a global perspective. American Institute for Cancer Research, Washington DC

Yusuf S, Collins R, Peto R (1984) Why do we need some large, simple randomized trials? Stat Med 3:409–422

Cluster Randomized Trials 10

Michael J. Campbell

Contents

M.J. Campbell
Medical Statistics Group, Health Services Research, ScHARR, University of Sheffield,
Sheffield, UK

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_47,

10.1 Introduction

10.1.1 Basics

A cluster randomized trial is one in which groups of subjects are randomized rather than individuals. They are sometimes known as group randomized trials. This chapter will describe the design and analysis of such trials. Examples of cluster trials in health are given in Box 10.1.

Cluster trials are used widely in the evaluation of interventions in health services research. They can be divided into two types. The first type is exemplified by the first three rows in Box 10.1: community randomized trials where the clusters are complete communities (some authors call these "large field trials"). These are generally characterized by a relatively small number of clusters each enrolling a large number of subjects. The aim of the trial by Grosskurth et al. (1995) cited in Hayes and Moulton (2009) was to reduce the prevalence of HIV infection by treating other sexually transmitted diseases. It involved six intervention communities and six matched control communities. In each, a random sample of 1,000 adults was selected in each community and followed up for 2 years to measure the incidence of HIV infection. The trial COMMIT (Gail et al. 1992) was to test an intervention aimed at communities to encourage citizens to stop smoking. It had 11 matched pair clusters. The sex-education trial by Wight et al. (2002) randomized 25 schools, 13 into intervention and 12 control, and interviewed all 13- to 14-year-olds at the schools and the same children after 2 years.

The second type of cluster trial is closer in design to an individually randomized trial. It typically uses more clusters and relatively smaller cluster sizes. Examples of "small cluster size" trials are given in the second half of Box 10.1. As an example of the second type, consider in more detail the DESMOND trial described by Davies et al. (2008) (DESMOND – Diabetes Education and Self Management Ongoing and Newly Diagnosed), which involved 105 general practices in the intervention and 102 in the control. The purpose of the trial was to investigate whether an intensive education package can be used to reduce glycosylated haemoglobin (HbA1c%) in patients who have type II diabetes. In the UK diabetes is usually treated in primary

Box 10.1. Examples of cluster trials

Unit	Intervention	Example
Rural communities	Treatment of coexisting disease	Grosskurth et al. (1995)
Communities	Education	Gail et al. (1992)
Schools	Education packages	Wight et al. (2002)
Groups	Diabetes education	Davies et al. (2008)
Doctors	Patient-centered care	Kinmonth et al. (1998)
Patients	Teeth fillings	Soncini et al. (2007)

care, and it was deemed impossible to randomize people in the same practice to different treatments. Thus practices were chosen (at random) as either “intervention” practices or “control” practices. DESMOND is usually taught as a course to groups of eight people at the same time, so the course was the cluster in this case. Kinmonth et al. (1998) randomized general practitioners into those who would receive training in “patient-centered care” and those who did not. A total of 21 practitioners were trained and 20 acted as controls. It would be difficult or impossible for a doctor to change from “patient-centered care” to “paternalistic” care with successive patients. The outcome was measured by HbA1c% in their diabetic patients. Soncini et al. (2007) looked at the survival of amalgam versus composite fillings in teeth and randomized 267 children into each group. It was deemed simpler to ensure each child either had amalgam or composite fillings and so survival times of the fillings will be clustered by mouth.

The main reason for using a cluster trial is fear of contamination. This occurs when subjects in the control group are exposed to the intervention. Thus people living in the same community could not fail to notice a mass education program delivered on the television or local newspaper. In the DESMOND trial, patients may wonder why people with the same doctor were getting different treatment and demand the same for themselves. Doctors trained in a new technique will find it difficult to revert to an old technique at the toss of a coin and so may not deliver the standard treatment as they used to do before being trained to deliver the new treatment. Another reason is that it may be more effective or cheaper to deliver an intervention to a group. For example, patients in the same education program will interact with each other and may learn more than if learning on their own. This was particularly true with DESMOND, where patients learned from each other as well as the trainer. A third reason for adopting a cluster design is administrative convenience or necessity; it is often easier to deliver an intervention to a group of people when it may involve an expensive piece of equipment or training health professionals or it may be impossible to randomize individuals. Again, this was true of DESMOND, where it was much cheaper to deliver the intervention in groups. Sometimes it appears easier to get ethical consent when all of a group are getting the intervention.

The most important point, with regard to the analysis, is that observations are not independent. Observations within a particular cluster are correlated, and although this correlation may be weak, it can have a major effect on the analysis as we shall see.

It is worth defining a few terms. The *intraclass correlation (ICC)* is the ratio of the between cluster variance to the total variance of an outcome variable and is often denoted by ρ. Different designs can lead to different formulas for estimating ρ. A simple method is described in the next section. The *design effect (DE)* is the ratio of the variance of an outcome measure when clustering is accounted for to the variance of the outcome measure when clustering is not accounted for. It is often referred to as the *variance inflation factor (VIF)* since it measures the amount that one should increase a variance estimate obtained by ignoring clustering to allow for the clustering effect. For clusters of equal size m, it can be shown that

$DE = 1 + (m - 1)\rho$. This is also called the *sample size inflation factor (SSIF)* in chapter ▶Generalized Estimating Equations of this handbook, since the same factor that inflates the variance will also inflate the required sample size. Extensions of this formula to the case of variable cluster sizes are given in Sect. 10.2.2.3.

10.1.2 Looking at Data

Figure 10.1 shows the HbA1c% in diabetic patients after 1 year from randomization by practice and by intervention/control from the DESMOND trial (Davies et al. 2008).

We are interested in the difference in the mean HbA1c% for intervention and control. However, one can see that there is a good deal of variation within practices but that some practices have in general high values and some practices have low values. This illustrates the key point: that we cannot think of the outcomes for individuals as being independent, we need to allow for the fact that two people in the same practice are more similar than two people selected at random from different practices. The intraclass correlation ρ is a measure of how much subjects within a cluster are correlated. It is the ratio of the between cluster component of variance σ_B^2 to the total variance $\sigma_B^2 + \sigma_W^2$ where σ_W^2 is the variance within clusters. We can estimate these using a simple analysis of variance. This is shown in Table 10.1 as the

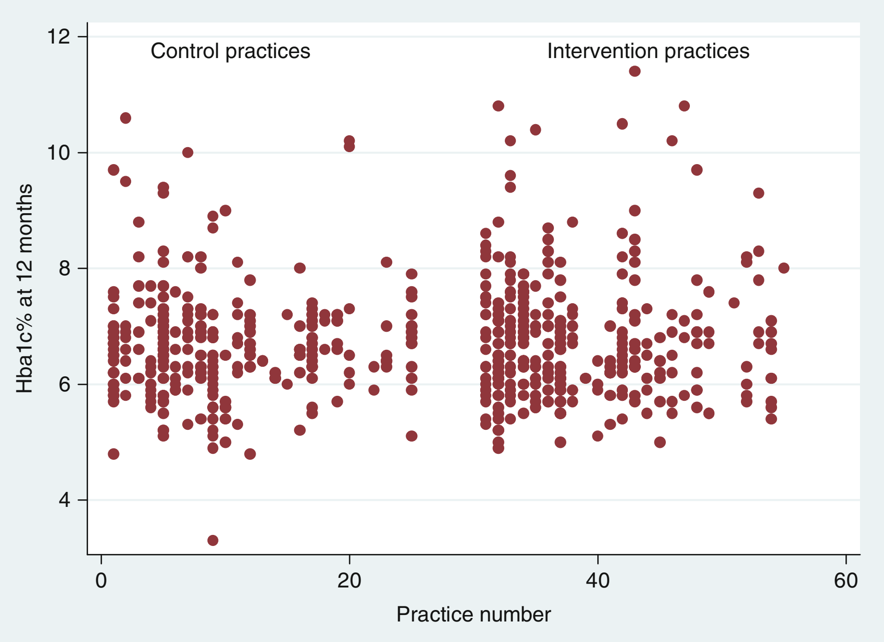


Fig. 10.1 HbA1c (%) at 1 year by control/intervention from DESMOND (Davies et al. 2008)

Table 10.1 Output showing ANOVA to estimate the ICC for data from DESMOND using Stata v11

```
Loneway hba1c12 practice
Source                  SS            df      MS              F       Prob > F
Between practice         77.357747     46     1.6816902       1.84    0.0009
Within practice         539.41716     589      .91581861
Total                   616.77491     635      .97129907

         Intra-class       Asy.
         correlation       S.E.        [95% Conf. Interval]

            0.05914      0.02832        0.00364      0.11464
```

output from the Stata command "loneway" (Statacorp 2009). We have 636 subjects in 47 practices. Here, σ_W^2, the *within practice mean square* (*MS*), is 0.9158. For a fixed cluster size, m, we can estimate the between cluster component of variance from the fact that the *between practice* $MS = \sigma_B^2 + m\sigma_W^2 = 1.6817$. Since here m is about 13.5, we find that $\sigma_B^2 = 0.0567$ and so ρ is $0.0567/(0.0567 + 0.9158) = 0.0583$. The program gives $\rho = 0.0591$, which is slightly different since it takes into account variable cluster size. It is important to appreciate that this procedure gives reasonable values for ρ even for binary outcomes, since it is a moment estimator and does not require distributional assumptions.

10.1.3 Overview of Chapter

Section 10.2 is concerned with the design of cluster randomized trials and how to estimate the number of patients and the number of clusters required. Section 10.3 discusses the analysis and presentation of such trials. Section 10.4 discusses other considerations for cluster trials and software for their analysis. Section 10.5 concludes the chapter and suggests further reading.

10.2 Design of Cluster Randomized Trials

10.2.1 Cohort Versus Cross-Sectional Designs

Many community intervention trials are longitudinal in nature, allowing a choice between a cohort design and a cross-sectional design. For a cohort design, clusters are randomly assigned to intervention groups, with or without stratification. Cohorts sampled from each cluster are then measured over two or more time points, with at least the first measurement occurring before randomization. They are useful in looking at how the intervention changes the health or behavior of individual subjects. The baseline and follow-up subjects are the same people. People who drop out will often be different from those that stay and so the follow-up group may not be typical of the whole population from which the cohort was chosen. Thus it is important to report drop-out rates and do sensitivity analyses to consider whether the nature of the drop-outs may affect the conclusions.

In contrast a cross-sectional design involves randomizing large groups of subjects, such as towns. A random sample is taken before and a random sample taken after the intervention with similar samples taken in the control population. Thus the subjects before and after the intervention are not necessarily the same. On the one hand, a cross-sectional study should be a representative sample of the population, since it is based on a random sample. On the other hand, some of the sample may have recently arrived in the population and so not received the intervention. This will reduce the size of the contrast between the intervention clusters and the control clusters. Cross-sectional designs are useful when the main focus is on change in behavior or health in a community. It can be helpful in some situations to ask those after the intervention whether they were aware of it. For example, in an evaluation of a government advertising campaign, we asked subjects after the campaign if in fact they had seen it (Mills et al. 1986). Only 31% of the sample were, in fact, aware of the campaign, which may partly explain its lack of effectiveness.

Because responses within the same subject often have a strong positive correlation, one can use the baseline measurement as a covariate and usually this will reduce the standard error of the treatment effect. Thus in theory a cohort design may be more efficient than a cross-sectional one. However, Feldman and McKinlay (1994) presented a unified statistical model that embraces both designs as special cases, thus allowing an assessment of how the values of different design parameters affect their relative precision. A principal conclusion from their investigation was that cohort designs have unique disadvantages that may outweigh any advantage in theoretical efficiency. The first of these is related to possible instability in cohorts of large size, with the resulting likelihood of subject loss to follow-up. Although this disadvantage can be compensated for by oversampling at baseline, this might well negate the original reasons for adopting a cohort design. Differential loss to follow-up by intervention group also creates the risk of bias. The second disadvantage is related to the issue of representativeness of the target population, which is invariably hampered by the aging of the cohort over time. Assuming that changes related to the aging process are independent of the intervention assignment, this effect will not invalidate the principal comparison of interest. However, it does imply that a difference observed in a cohort trial with respect to a given outcome variable cannot be directly compared to the corresponding difference between observed cross-sectional samples. Thus if the primary questions of interest focus on change at the community level rather than at the level of the individual, cohort samples are the less natural choice. This point was discussed by Ukoumunne and Thompson (2001) and by Nixon and Thompson (2003), who described and compared several approaches that might be taken to the analysis of repeated cross-sectional samples.

10.2.2 Power and Sample Size

10.2.2.1 Number of Clusters and Number of Subjects per Cluster

In cluster randomized trials, there are two sample size choices to be made: the number of clusters and the number of subjects per cluster. The usual situation is where the cluster size is fixed, and to increase the power we need to increase the number of clusters.

Suppose we needed n' subjects in an individually randomized trial to detect an effect size δ with two-sided significance α and power $1-\beta$ (e.g., using the tables in Machin et al. 2008). Then to allow for clustering where we have equal clusters of size m, we should increase the sample size (using *VIF/SSIF*) to n where n is given by

$$n = n'(1 + (m-1)\rho) \tag{10.1}$$

to achieve the same power and significance level (Hsieh 1988). The number of clusters is determined by $k = n/m$.

An alternative situation is where the number of clusters is fixed, and one wishes to determine the number of subjects per cluster. Since the design effect requires knowledge of the number of subjects per cluster, m, one has to guess m first, to find n and then recalculate m from $m = n/k$ and then reiterate. A simpler solution is to use the fact that (Campbell 2000)

$$m = \frac{m'(1-\rho)}{1-m'\rho}, \tag{10.2}$$

where $m' = n'/k$ is the number of subjects per cluster required before adjusting for clustering. Suppose that for a given effect size, significance level, and power, we require m' subjects per cluster in an individually randomized trial. If ρ is greater than $1/m'$, then m becomes negative, which is impossible and so one can never achieve the required power simply by increasing the number of subjects per cluster, and one will have to increase the number of clusters. Even if ρ is only slightly less than $1/m'$, the numbers per cluster become very large. Thus a useful rule of thumb for continuous outcomes is that the power does not increase appreciably once the number of subjects per cluster exceeds $1/\rho$. For example, if it is believed that the *ICC* is about 0.05, then it is not worth enrolling more than about 20 subjects per cluster for a continuous outcome. However, if the *ICC* of a continuous outcome is near 0.001, which is often typical of community intervention trials, then a sample of 1,000 subjects per cluster may be worthwhile, particularly if recruiting new clusters is difficult. Of course, with binary outcomes, when the incidence is low, large numbers of patients per cluster are required unless the effect of the intervention is very large.

Flynn et al. (2002) addressed the issue of whether it is worth recruiting an extra cluster, or to recruit more individuals to existing clusters. They showed how the use of contour graphs of power by number of clusters per treatment arm and cluster size can be usefully exploited. For example, consider a hypothetical trial in which

18 clusters have already been recruited in each of two treatment arms and in which at least 30 individuals can be recruited from each cluster. Suppose the *ICC* is 0.05 and the target standardized difference is 0.25. We then currently expect about 75% power. To achieve at least 80% power, we can show there are two options: (1) recruiting 20 extra individuals in each existing cluster; (2) recruiting two extra clusters in each arm. The question may then be which of these options is the least costly, and this would be the option to choose.

Although values of ρ in cluster randomized trials tend to be small (typically around 0.05 for primary care trials (Campbell 2000)) and in community randomized trials even smaller (usually less than 0.01 and often near 0.001, Donner and Klar 2000), the resulting inflation of the sample size may be very substantial when combined with clusters of large size. For example, in the trial of HIV reduction (Grosskurth et al. 1995), communities were on average 1,000 adults and so even an *ICC* as low as 0.0001 would have the effect of doubling the required sample size. Results from earlier studies in a specific setting of design effects likely to arise in cluster randomized trials implemented are also helpful to investigators. Gulliford et al. (2005) gave examples of variance components for some common outcomes which can be helpful for future planning.

The sample size formulas we give here assume that data for sample size estimation are obtained from a single sample of clusters from the population of interest, that is, that the intervention itself is not associated with the cluster size. The problem here is that variable cluster sizes will affect the power, and this will be discussed in Sect. 10.2.2.3. Examples where this is not true have been given by Campbell (2000). A particular example is Kinmonth et al. (1998) where the intervention was to train doctors in treating newly diagnosed diabetics, and these same doctors were the ones who diagnosed the diabetes and recruited the patients. They found the intervention clusters were larger than the control because the doctors who had received the intervention seemed more likely to find people with diabetes. This is known as recruitment bias and is a particular problem with cluster trial.

In summary, ignoring clustering effects in the design stage of a trial can lead to an elevated type 2 error, while ignoring it at the analysis stage inevitably leads to an elevated type 1 error. In other words, if an investigator ignores clustering when planning a study, the study is likely to be too small and so underpowered. If the investigator ignores clustering in the analysis, then the standard error of the estimate is likely to be underestimated, and so the observed p-value will be too low, and results declared significant at a given level, when in fact the null hypothesis should not have been rejected at that level.

10.2.2.2 Allowing for Imprecision in the *ICC*

Much of the sample size literature deals with the difficulty of obtaining accurate estimates of between community variation, and hence of ρ, that are needed for sample size planning.

In practice estimates of ρ for a given outcome variable are usually derived from previously reported studies using similar randomization units. However these estimates are frequently based on studies which themselves are of small size, and thus their inherent inaccuracy may lend the investigators a false sense of confidence.

Turner et al. (2004) have shown how to incorporate uncertainty in the *ICC* in a Bayesian framework to obtain an "average" power (cf. chapter ▸Bayesian Methods in Epidemiology of this handbook). They discussed the use of prior distributions for the ICC and showed how the uncertainty about this parameter can be expressed in the form of a parametric distribution which naturally leads to a distribution of projected power for any particular design. This Bayesian approach toward the determination of sample size could then be followed by a statistical analysis within the traditional frequentist framework. If the total sample size were fixed, and $n = m \times k$, it is better to increase the number of clusters k and have smaller cluster sizes. Increasing the number of clusters also considerably reduces the lower limit of the posterior distribution of power. In other words, uncertainty in the *ICC* will produce uncertainty in the actual power of a study, but a design based on a greater number of clusters has less chance of having a very low power. In general, one should try and recruit as many clusters as possible.

An alternative approach to dealing with uncertainty in observed values of ρ is described by Feng and Grizzle (1992), who proposed the use of a method similar in principle to the bootstrap procedure. For a simple discussion of the bootstrap in this context, see Carpenter and Bithell (2002). Their approach requires the simulation of results of studies of the same size to that which yielded the observed estimate. One then can substitute the values of ρ obtained from each simulation into the appropriate sample size formula to generate a distribution of projected powers, followed by the selection of a point on this distribution, for example, the 90th percentile, that reflects the degree of conservativeness desired.

10.2.2.3 Allowing for Varying Cluster Sizes

Variation in cluster size is another source of imprecision, and Kerry and Bland (2001) following Donner et al. (1981) suggested using

$$DE = 1 + (m_a - 1)\rho, \text{ where } m_a = \sum m_i^2 / \sum m_i. \tag{10.3}$$

The problem of using this formula is that the individual cluster sizes must be known prior to conducting the trial.

Eldridge et al. (2006) modified (10.3) to give

$$DE = 1 + ((cv^2 \frac{(k-1)}{k} + 1)\bar{m} - 1)\rho, \tag{10.4}$$

where $\bar{m}$ is the mean cluster size, s_m is the standard deviation ($s.d.$) of the cluster size given by $s_m = \sqrt{\sum (m_i - \bar{m})^2 / (k-1)}$, and the coefficient of variation $cv = s_m / \bar{m}$. They suggested that formula (10.4) is more practical than (10.3) since the value of cv may often be known in advance. Eldridge et al. (2006) also provided some examples of the coefficient of variation for sample sizes typically seen in cluster randomized trials. Data from similar trials would be the first source of values for cv. Alternatively one could ask what is likely to be the maximum and minimum size cluster and estimate s_m as the range is divided by 4 (although strictly speaking this would require the data to be normally distributed).

Eldridge et al. (2006) showed empirically that as the average cluster size increases, the coefficient of variation tends toward 0.65. Primary care trials in the UK tend to have values of cv between 0.42 and 0.75. In the trial of Davies et al. (2008), the mean cluster size was about 14, and the $s.d.$ of cluster size was 12, suggesting a cv of about 0.86 which is somewhat more variable than many. The range of cluster sizes was 1 to 48, so the rule of thumb of estimating s_m as $(48 - 1)/4 = 11.75$ is quite accurate. We give an example of a sample size calculation in the next section.

10.2.2.4 Example of Effect of Variable Cluster Size

Suppose an investigator stated that the expected number of patients per cluster was 10 and the estimated intra-cluster correlation coefficient was 0.05. A preliminary sample size calculation showed that the estimated sample size required for a given power, significance level, and effect size without taking account of clustering was 200 patients and so 20 clusters. Then the estimated sample size required taking account of clustering but ignoring variation in cluster size is

$$n = (1 + (\overline{m} - 1)\rho)x\,200 = 1.45 \times 200 = 290 \text{ patients } = 29 \text{ clusters.}$$

We can argue that a conservative estimate of the expected minimum size of a cluster is 1 patient (no cluster can have less than 1) and the expected maximum is 30 (since we would stop recruiting above this level). The cv is then estimated $(29/4)/10 = 0.725$. From Eq. 10.4, the design effect

$$DE = 1+((1+cv^2\times 28/29)\overline{m}-1)\rho = 1+((1+0.725^2\times 0.966)10-1)0.05 = 1.70.$$

Thus we would need $1.70 \times 200 = 340$ patients $= 34$ clusters, an increase of 5 clusters to allow for cluster size variable.

10.2.2.5 Sample Size Re-Estimation

When we do not know the value of the variance components required to ascertain a sample size, one suggestion is to conduct a pilot study to provide these estimates (Friede and Kieser 2006). We can then obtain an estimate of how many more patients we will need to recruit and use the patients from the pilot in the final analysis (an internal pilot). This procedure has been extended to cluster randomized trials by Lake et al. (2002). With this approach, we conduct a pilot cluster trial, and at an interim point in the study, several nuisance parameters, including ρ, the mean cluster size and measures of cluster size, variation, are estimated, followed by re-estimation of the final required trial size. Although this procedure is most suited to trials that randomize a relatively large number of clusters, such as families or households, over an extended period of time, Lake et al. (2002) pointed out that it could also be applied to at least some community intervention trials provided the participating clusters are recruited prospectively. However, Turner et al. (2004) showed that imprecision in the estimate of ρ is not accounted for in this application and suggested their Bayesian method could easily be extended to do so. More experience on the application of internal pilot studies to such trials is clearly needed.

10.2.2.6 Adjusting for Covariates

Sometimes the reason for a positive *ICC* is that subjects within a cluster have similar covariates. Several authors have shown that adjusting for covariates either at the community or individual level can improve the power of a trial by reducing the magnitude of the between cluster variation (Campbell 2000; Feng et al. 1999). Additional gains in power may be realized by modeling individual level covariates (e.g., age) and also cluster level covariates (e.g., mean cluster age) as described by Klar and Darlington (2004). Moerbeek (2006) suggested that often a cheaper strategy than recruiting another cluster would be to measure additional covariates related to the outcome in order to reduce the variance. This, of course, requires the cost of measuring the covariate to be relatively small and the correlation of the covariate and the outcome to be reasonably high.

10.2.3 Matched Pair Trials

10.2.3.1 Design of Matched Pair Studies

Because cluster trials involve randomizing relatively low numbers of groups, we cannot rely on randomization to ensure balance between treatment arms in important prognostic variables. A common technique to try and ensure balance is to match clusters into pairs and then randomly allocate one member of each pair to the intervention and one to the control.

Matched pair studies are not frequently seen in clinical trials randomizing individual subjects to different intervention groups. However, they have proven to be the design of choice for many investigators embarking on a community intervention trial largely because of the perceived ability of this design to create intervention groups that are comparable at baseline with respect to important prognostic factors, including, for example, community size and geographical area. The relatively small number of communities that can be enrolled in such studies further enhances the attractiveness of pair matching as a method of reducing the probability of substantively important imbalances that may detract from the credibility of the reported results.

Freedman et al. (1990) investigated the gain in efficiency obtained from matching in a community intervention trial. This was done in the context of the COMMIT trial (Gail et al. 1992). Eleven pairs of communities were matched on the basis of several factors expected to be related to the smoking quit rates, such as community size, geographical proximity, and demographic profile. Within each matched pair, one community was allocated at random to the intervention group, with the other acting as the control. It is also interesting to note that this trial was one of the first large-scale community intervention studies to use formal power considerations at the planning stage and, perhaps not coincidentally, to be substantially larger in size than its predecessors.

The gain in efficiency (measured by the sample size required for a given power and effect size) due to matching may be quantified by the factor $G = 1/(1 - \rho_{\mathrm{m}})$, where ρ_m is the correlation between members of a pair with respect to the outcome variable. This latter quantity is simply the Pearson correlation between outcomes for the intervention and control. Thus if the correlation was 0, there would be no gain in efficiency, whereas a value of 0.5 would reduce the required sample size by 50% if matching were employed. Freedman et al. (1990) showed that matching can lead to considerable gains in statistical precision when it is based on an effective surrogate for outcome. However, since G is simply the ratio of population variances ignoring or accounting for pair-matching, it does not take into account the difference in degrees of freedom for estimating these variances, a factor which is particularly relevant in trials enrolling a small number of communities. For example, in the COMMIT study, there were 11 matched pairs. The degrees of freedom associated with the error from the paired differences in event rates would be only ten, as compared to the 20 degrees of freedom available for an unmatched analysis. This issue was subsequently addressed in detail by Martin et al. (1993), who concluded that for studies having no more than 20 pairs, matching should be used for the purpose of increasing power only if the investigators are confident that ρ_m exceeds 0.20. By considering the practical difficulties that often arise in securing "good" matches, they also concluded more generally that "matching may be overused as a design tool" in community intervention trials.

These considerations suggest that a tempting strategy in practice may be to perform an unmatched analysis of data arising from a matched pair design, particularly when matching is adopted mainly for the purpose of avoiding a "bad" randomization. The effectiveness of such a strategy was investigated by Diehr et al. (1995), who concluded on the basis of an extensive simulation study that breaking the matches can actually result in an increase in power when the number of pairs is less than ten. Thus the loss in precision identified by Martin et al. (1993) in the presence of weak matching correlations can be at least partially regained.

A secondary objective of many community intervention trials is to investigate the effect of individual level risk factors on one or more outcome variables. Focusing on the case of a continuous outcome variable, Donner et al. (2007) showed that the practice of performing an unmatched analysis on data arising from a pair-matched design can lead to bias in the estimated regression coefficient and a corresponding test of significance which is overly liberal. However, for large-scale community intervention trials, which typically recruit a relatively small number of large clusters, such an analysis will generally be both valid and efficient.

10.2.3.2 Limitations of Matched Pairs Designs

Klar and Donner (1997) explored some further limitations of the matched pair design that are more general in nature. These limitations arise largely from the total confounding of the intervention effect with the natural variation that exists between two members of a matched pair. One consequence of such confounding is that it precludes the use of standard methods for estimating the underlying *ICC*, which in turn reduces analytical flexibility. For example, a secondary objective of many studies is to estimate the effect of selected individual level risk factors on one or more outcome variables using regression modeling. However, the calculation of appropriate standard errors for the regression coefficients obtained from such a model requires a valid estimator of ρ. Thus, although it is possible to perform adjustments for the effect of such risk factors, the task of testing for their independent relationship with outcome is more difficult (Donner et al. 2007). It is difficult to directly model the joint effects of cluster level and individual level risk factors, and the matched pair design frequently does not bring large gains in precision. Klar and Donner (1997) recommended that greater attention should be paid to the possibility of adopting a stratified design, in which two or more clusters are randomized to each intervention group within strata. This design may be particularly attractive when investigators find it challenging to create matched pairs that correspond to unique estimates of risk for each pair. Most importantly, the cluster level replication inherent in this less rigid allocation scheme removes many of the analytical limitations associated with pair-matching, while increasing the degrees of freedom available for estimating error.

Perhaps the most commonly adopted matching factors in large-scale community randomized trials have been cluster size and geographical area (e.g., urban vs. rural). Matching by cluster size is attractive not only because it protects against large imbalances in the number of subjects per intervention group, an efficiency consideration, but also because cluster size may be associated with other important but unaccounted for baseline variables, such as socioeconomic status and access to health-care resources (Lewsey 2004). Matching by categorized levels of baseline versions of the trial outcome rate would also seem attractive. However, results reported by Feng et al. (1999) suggest that if the primary interest is in change from baseline, such matching is not likely to add benefits in power beyond that yielded by an analysis of change scores. This is because including the baseline in the model analysis is as effective as matching for baseline.

10.2.4 Problems with Identifying and Recruiting Patients to Cluster Trials

In the trial conducted by Kinmonth et al. (1998), the subjects were newly diagnosed people with type II diabetes. The doctors who recruited them were the same doctors who were given training in patient-centered care. After the trial, it was discovered that there were more patients diagnosed in the intervention arm than in the control possibly because the doctors were unblind to treatment. However, concealment of allocation is usually regarded as crucial for individually randomized trials, and one of the advantages of randomization is that it ensures that it is impossible to predict which treatment the next potential recruit will get (see also chapter ▶Clinical Epidemiology and Evidence-Based Health Care of this handbook). This advantage is lost for cluster trials where randomization of clusters is usually accomplished at the start of the trial and so concealment is more difficult. In most cases, it is impossible to conceal the identity of the treatment from the patients when they receive it but it is useful to conceal what treatment patients will get until after they are recruited to the trial. In a new trial, currently in the planning stage, an insulin pump is being tested in patients with type II diabetes. The patients are educated in the use of the pump in groups of size six. The patients are recruited to the trial and asked to give consent to either treatment. When six have been recruited, they are randomized to either the pump therapy or control. In this way the recruiters are ignorant of the treatment the patients will receive. Eldridge et al. (2010) also discussed various options for trying to ensure concealment. These include recruiting clusters and patients before randomization, masking recruiters, or using a standardized recruitment procedure across clusters to try and ensure the procedure was not affected by subsequent treatment.

10.3 Analysis of Cluster Randomized Trials

10.3.1 Cluster Specific Versus Marginal Models

Assume that the clusters are sampled from a larger population and the effect of any particular cluster i is to add a random effect Z_i to the outcome Y. We assume the Z_is have the same distribution for all i. We add a covariate X for the treatment effect, where $X = 1$ for the intervention and $X = 0$ for the control. Suppose the effect of an intervention is to add an amount β_1. We assume that the actual cluster effect is a separate and independent effect to that of the treatment effect. A *cluster specific (CS)* model measures the effect on Y of changing X, while Z is held constant. This is a common model for longitudinal data, where it is possible to imagine, say in a cross-over trial, a treatment value changing over time. A suitable model might be

$$E(Y_i|Z_i) = \beta_0 + \beta_1 X + Z_i, \tag{10.5}$$

where $E(Y_i|Z_i)$ is the expected value of the outcome conditional on Z_i. We further assume that $E(Z_i) = 0$ and $\mathrm{var}(Z_i) = \sigma_Z^2$ and the Z_is are independent of the fixed effect X for all i. The distribution of Z_i is generally assumed to be normal, but for binary data, a gamma distribution can be used. In a Bayesian context, other distributions such as a t-distribution can be also used (see Sect. 10.3.4).

Equation 10.5 can be generalized to any outcome variable Y_i (continuous or binary), with expected value μ_i and a generalized link of the form

$$g(\mu_i) = \beta_0 + \beta_1 X + Z_i, \tag{10.6}$$

where the function g is assumed strictly monotone and differentiable.

However, in a cluster randomized trial, everyone in a cluster receives the same treatment, and although a *CS* model can be fitted, the result can be interpreted only theoretically. There is an analogy here with the "counterfactual" argument for causality in epidemiology (see also chapter ▸Basic Concepts of this handbook), where we interpret casual effects as being the difference in outcome from either exposure to a hazard or non-exposure in the same person, even though in practice this is not observed.

An alternative method is to fit a model which looks at the average effect of X over the range of Z. This is the so-called *population averaged (PA)* or *marginal* model. Consider a model where we fit only X and ignore Z_i, so that

$$g(\mu_i) = \beta_0^* + \beta_1^* X. \tag{10.7}$$

Model (10.7) is a *PA* model, that is, we estimate the effect of X on Y as averaged over all the clusters i.

Neuhaus and Jewell (1993) contrasted the approach between cluster specific models and population averaged models by observing that Model (10.7) is simply Model (10.6) with the variable Z omitted. If we assume that the coefficients for X in the two models are related by $\beta_1^* \approx B\beta$, where B is the bias factor, then they show that for a linear model and a log-linear model $B = 1$, and so the interpretation of cluster specific and marginal models is the same. However for a logistic link, $\mu =$ logit $(P(Y = 1|X, Z))$, they showed that $B \approx 1 - \rho$ where ρ is the correlation of the Ys within clusters assuming $\beta_1 = 0$. Since $0 < \rho < 1$, so $0 < B < 1$, and so the general effect of using a population averaged model is to attenuate the regression coefficient toward zero. One can also see that for a logistic link, the greater the variability of the random variable Z_i, and so the greater the intracluster correlation, the greater the attenuation. However, as discussed earlier, the value of ρ in community randomized trials is usually less than 0.01, and this suggests that the bias in assuming a marginal model should not be great.

Since $B = 1$ for a log link, this would suggest that in prospective studies such as clinical trials, it would be advantageous to use a log-linear model which estimates the relative risk rather than a logistic model which yields odds ratios (Campbell 2008). However, experience has shown that in general logistic models are easier to fit and have fewer convergence problems. These can arise with a log-linear model when the fitted values for the risk become greater than 1 or less than 0. This is more likely to happen when the number of events is relatively high.

10.3.2 Standard Methods of Analysis

10.3.2.1 Inflating the Standard Error

For a linear model, we assume the observed outcome y_{ij} is the outcome of the random variable Y_{ij} for the jth subject ($j = 1, \ldots, m_i$) in the ith cluster ($i = 1, \ldots, k$), and it differs from the expected value by a random error. We write

$$Y_{ij} = \beta_0 + \beta_1 X_i + Z_i + \varepsilon_{ij} \tag{10.8}$$

and we assume $E(\varepsilon_{ij}) = E(Z_i) = 0$, ε_{ij} and Z_i are independent, var$(\varepsilon_{ij}) = \sigma^2$, and var$(Z_i) = \sigma_Z^2$. In a trial with no other covariates, X_i is an intervention indicator variable (0, 1) which depends only on whether the cluster i is in the intervention group or not. An *exchangeable* correlation structure is assumed, which effectively means that one can exchange subjects j and j' within a cluster without changing the covariance. This breaks down if subjects are measured more than once (e.g., at baseline and at follow-up) since the correlation of the same subject measured twice will not be the same as the correlation of two different subjects within a cluster. It also means that the *ICC* must be assumed to be the same within each arm of the trial, an assumption which is guaranteed in a randomized trial under the null hypothesis of no intervention effect, but may not be true under the alternative hypothesis that the intervention affects the outcome.

Let $\bar{d}$ be the estimate of the difference in means between the intervention and control group and suppose there are $k/2$ clusters in each group (k assumed even). Then we can show that

$$\operatorname{var}(\bar{d}) = \frac{4\sigma^2}{mk} + \frac{4\sigma_Z^2}{k}. \tag{10.9}$$

The first term in (10.9) is simply the variance that would have been obtained if the data were not clustered. Equation 10.9 can be rewritten as

$$\operatorname{var}(\bar{d}) = 4(\sigma_Z^2 + \sigma^2)VIF/mk,$$

and we can estimate $\sigma_Z^2 + \sigma^2$ by the pooled variance of the outcome variable over groups. Thus a technique originating in sample survey is to simply multiply the variance obtained from ignoring the clustering by the variance inflation factor *VIF*. The design effect given in Sect. 10.1 may be estimated by replacing the *ICC* with its sample estimate (Donner and Klar 2000). A simple test of whether a parameter is zero, known as a modified Wald test, is to divide an estimate of the parameter by its modified standard error which is then compared to the quantile of a standard normal distribution. The authors give a number of methods for continuous and binary outcomes which modify the standard error associated with either the t-test or the chi-squared test respectively. It is important to note that the estimate of the treatment effect is unchanged, only the standard error is inflated. An alternative method is to use the so-called "sandwich," "robust" or Huber-White estimator (Huber and Ronchetti 2009) which has a long history in econometrics for estimators with continuous data and with heterogeneous variances. The advantage of the robust standard error is that one does not need to estimate the *ICC* separately before conducting the analysis.

10.3.2.2 Summary Measures

A simple method, which is applicable to both binary and continuous data, is the method of summary measures, as popularized by Matthews et al. (1990). For continuous data, one uses the mean of each cluster, and for binary data, one would use the proportion of events (or a transformation such as the logit). This works best when the clusters are all approximately the same size. It gives equal weight to each cluster, irrespective of size, and is a cluster specific method. It has a great deal to recommend since it simply uses the summary statistics for each cluster and is easy to apply without specialist software. However, one cannot adjust for individual level covariates directly using this approach.

10.3.2.3 Generalized Estimating Equations

The generalized estimating equations (GEE) method, developed by Liang and Zeger (1986) in the context of longitudinal studies, has proved to be very popular for the analysis of data arising from cluster randomized trials. It fits the *PA* model and uses an iteratively reweighting algorithm to estimate the parameters and a robust method (the "sandwich" estimator) for the standard error. It is described in more detail in chapter ▶Generalized Estimating Equations of this handbook. Use of the GEE yields a "shrinkage" estimator which is a compromise between no weighting and weighting by the sample size. It deals with the correlation within clusters by assuming a "working" correlation and then adjusting it according to the data. In cluster trials the choice is between an independent error structure and an exchangeable error structure. An independent error structure is plausible if in fact the intracluster correlation coefficient of the outcome variable is close to zero. An exchangeable error structure means that one can exchange subjects within a cluster and not change the correlation matrix. An exchangeable correlation structure effectively weights each mean by $m_i/(m_i\sigma_z^2 + \sigma^2)$. This weights the means by the sample size m_i when $\sigma_Z^2 = 0$ and gives equal weight when $\sigma^2 = 0$ or when the cluster sizes are all the same. In practice, estimates of the variance components, s_Z^2 and s^2, are used and so s^2 will always be greater than zero which implies the weight will vary unless the outcome were constant.

The use of robust standard errors means that even if a model has an incorrect variance-covariance structure, valid inferences can still be made. For example, one could have a model with an independent error structure and use robust standard errors. This is the same as using the variance inflated method described in Sect. 10.3.2.

GEE is used widely for hypothesis testing and confidence interval construction because it can control for the influence of potential confounders on outcome without the need to specify an underlying distribution for the sample observations. The robust variance estimation relies on between cluster information to assure the validity of the resulting inferences. It is therefore important to be wary of this approach to community intervention trials where the amount of such information tends to be relatively small.

Feng et al. (1996) recommended for continuous data that GEE should not be applied to trials having 20 or fewer clusters. It has been found that using a t-distribution (with a Satterthwaite type correction for the degrees of freedom to allow for unequal variances) and a technique known as the jackknife (Efron and Tibshirani 1998) improves the estimate of the standard error (Mancl and DeRouen 2001). Pan and Wall (2002) proposed replacing the GEE Wald test by approximate t- or F-tests. Although the proposed procedures showed type 1 errors closer to nominal than the usual Wald test, they were shown to be strictly applicable only in clusters of small size. It is therefore clear that more research is needed on the development of adjusted GEE procedures that can be applied to clusters of the size that typically arise in community intervention trials.

10.3.2.4 Random Effect Models

The alternative method of analyzing data from cluster trials is to use a cluster specific model (10.5 and 10.8). We now have to assume distributions for the two error terms. For continuous data the subject level error is assumed normal and for binary data it is assumed binomial. For continuous data the cluster level error is usually assumed normal, and also for binary data, although sometimes a gamma distribution is used. Although this model does not directly reflect the design of a cluster trial since treatment is contrasted to control within the same cluster, as stated earlier it does provide a valid estimate of the treatment effect. These models are also known as "mixed" models (since they contain a mixture of random and fixed effects) or "hierarchical" models since one can think of a hierarchy of clusters and then subjects nested within clusters.

The probability density of an observation from Eq. 10.8 conditional on Z_i is normal with mean $\beta_1 X_i$. Thus $P(y_{ij}|Z_i) = f(y_{ij}|Z_i,\beta,\sigma^2)$ where $f(\cdot)$ is the normal density function. Within a cluster and conditional on Z_i, we assume the observations are independent and so, given observations $y_{i1}, y_{i2}, \ldots, y_{im_i}$

$$P(y_{i1}, y_{i2}, \ldots, y_{im_i}|Z_i) = \prod_{j=1}^{m_i} f(y_{ij}|Z_i,\beta,\sigma^2).$$

This depends on the random variable Z_i, and to find the expected value, we integrate over possible values of Z_i to get

$$P(y_{i1}, y_{i2}, \ldots, y_{im_i}) = \int_{-\infty}^{+\infty} f(Z_i,\sigma_Z^2) \prod_{j=1}^{m_i} f(y_{ij}|Z_i,\beta,\sigma^2)dZ_i.$$

The full likelihood is the product of the above integrals over k clusters

$$L(\beta,\sigma_Z^2,\sigma^2) = \prod_{i=1}^{k} \int_{-\infty}^{+\infty} f(Z_i,\sigma_Z^2) \prod_{j=1}^{m_i} f(y_{ij}|Z_i,\beta,\sigma^2)dZ_i. \tag{10.10}$$

As discussed in Sect. 10.3.1, binary models using different link functions can estimate different population parameters and so deserve special consideration.

Let Y_{ij} (0 or 1) be the jth observation ($j = 1,\ldots,m_i$) in the ith cluster ($i = 1, 2, \ldots, k$). The cluster specific logistic model, following Eq. 10.6, is

$$\text{logit}\,(\pi_{ij}) = \beta_0 + \beta_1 X_i + Z_i, \tag{10.11}$$

where Z_i is the effect of being in cluster i and where $\pi_{ij} = E(Y_{ij}|X_i, Z_i)$. This model can be extended to include individual specific covariates X_{ij}. The random variable Z_i is assumed to be independent of X_i and may be usually assumed to be normally distributed with mean 0 and variance σ_Z^2 although sometimes a gamma distribution fits the data better. Given the Z_i, the Y_{ij}s are assumed independently distributed with binomial parameter π_{ij}.

The full likelihood L is given by

$$L(\beta, \sigma_z^2) = \prod_{i=1}^{k} \int \prod_{j=1}^{m_i} \pi_{ij}(\beta, Z_i)^{y_{ij}} \{1 - \pi_{ij}(\beta, Z_i)\}^{1-y_{ij}} f(Z_i, \sigma_Z^2) dZ_i. \quad (10.12)$$

Equation 10.10 can be solved directly to maximize the likelihood with respect to the parameters β, σ_Z^2, and σ^2 but not so Eq. 10.12. An early method for binary outcomes and which avoided the integration is a penalized quasi-likelihood approach, using a Laplace method for approximating the integral (Breslow and Clayton 1993). However, this has been replaced by methods which conduct the integration directly using Gaussian quadrature or other numerical methods to obtain estimates of the regression coefficients. Other methods, using iteratively generalized least squares (IGLS), are commonly used for hierarchical models (Goldstein 2002) and are implemented in the package MlWin.

10.3.3 Examples

10.3.3.1 The Analysis of Continuous Data

Table 10.2 gives the results from the DESMOND study (Davies et al. 2008) of the analysis of the outcome HbA1c%, which is treated as a continuous variable. The first row is the result of using a simple t-test on the means of the clusters. This ignores the size of each cluster. The second row uses a robust correction factor for the standard error. The estimate 0.0792 is what one would get from an analysis ignoring clustering, but the standard error is inflated using a "sandwich" estimator (cf. chapter ▸Generalized Estimating Equations of this handbook). The third row uses an exchangeable error structure and shows the effect of "shrinking" the smaller clusters toward the center. The random effects model using maximum likelihood gives nearly the same outcome as the GEE with exchangeable errors, a common finding. One can see that the GEE (independent errors) gives a smaller p-value than the other methods possibly because the independence assumption is implausible.

It is sensible to plot the residuals from the random effect to check for approximate normality. Figure 10.2 shows a plot known as a QQ plot, which plots the residuals against the value that would have been expected from a normal distribution with

Table 10.2 Results of analysis of continuous outcome HbA1c% from DESMOND

	Treatment effect	Std. Err.	z	$P > z$
t-test using means	0.0615	0.1321	0.47	0.64
GEE (independent errors)	0.0792	0.1118	0.71	0.48
GEE (exchangeable errors)	0.0531	0.1098	0.48	0.63
Random effects (max.lik.)	0.0518	0.1100	0.47	0.64

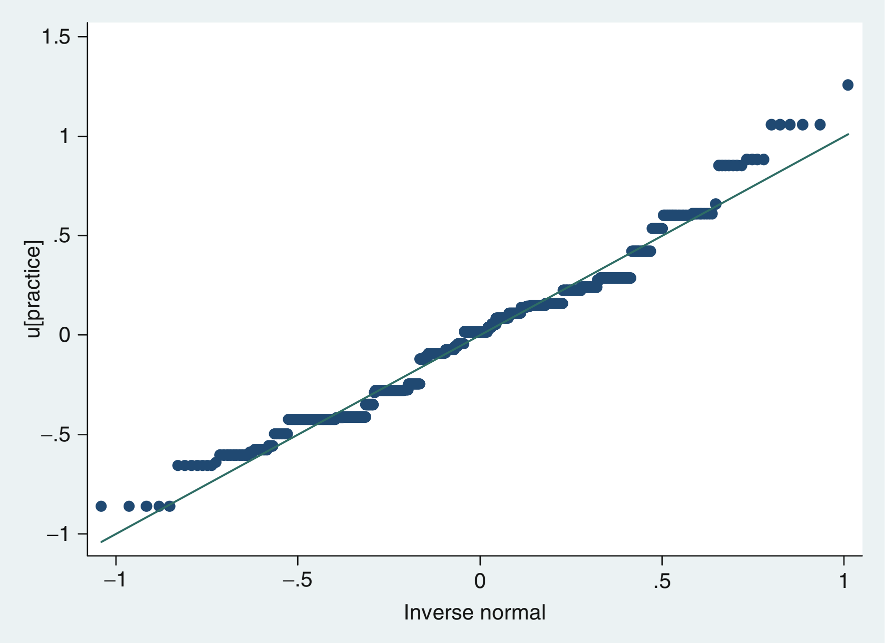


Fig. 10.2 A QQplot of the cluster level residuals from a random effect model from the DESMOND study

the same mean and standard deviation. If the residuals are normally distributed, one would expect this plot to be a straight line. The model fitting procedure means that the residuals are closer to a normal distribution than if we had not estimated the parameters from the data, but the plot is useful for gross departures from expected. The figure suggests that the residuals are plausibly normal. If the plot had been grossly away from normal, one would look for outliers, include potential covariates, and try transformations of the outcome to search for a better fitting model.

10.3.3.2 The Analysis of Binary Data

Table 10.3 shows the outcomes of three analyses from the DESMOND study (Davies et al. 2008) where the outcome is a binary indicator variable (HbA1c% > 7.5%). One can see that the GEE (population averaged) model with exchangeable errors gives a smaller odds ratio (*OR*) than that for independent errors. As with the continuous data analysis of Sect. 10.3.3.1, this is because smaller clusters are shrunk closer to the overall mean in a model with exchangeable errors and it is more likely that small clusters will have larger effects, the effect of which is diminished with exchangeable errors. The *OR* estimated using a random effects model is similar to that for the GEE population averaged model with exchangeable errors since the *ICC* is relatively small (0.061 in this case), but the GEE estimate is very slightly smaller as might be expected. Further discussion of these points has been given by Ukoumunne et al. (2007, 2008).

Table 10.3 Results of analysis of binary outcome for DESMOND using a logistic link

	Odds ratio for treatment	Std. Err.	z	P > z
GEE (independent errors)	1.8318	0.4689	2.36	0.018
GEE (exchangeable errors)	1.8056	0.4568	2.34	0.20
Random effects (adap. quad.)	1.8190	0.5012	2.17	0.30

Tests for the assumptions concerning the residuals with binary data are more difficult to achieve than for continuous data. They are most easily accomplished using Bayesian methods described in the next section. In view of the close agreement between the GEE (exchangeable) and random effects models, we will not pursue this further.

10.3.4 Bayesian Methods

An alternative method to solve Eqs. 10.10 and 10.12 is to use simulation via Markov Chain Monte Carlo (MCMC) algorithms. These are usually associated with a Bayesian analysis, but choice of suitable non-informative priors will yield results similar to the conventional likelihood methods. Spiegelhalter (2001) described methods for the Bayesian analysis of cluster randomized trials with a continuous response. This was extended to a binary outcome by Turner et al. (2001). They used Eq. 10.11 and looked at different prior distributions for σ_Z^2. Since σ_Z^2 is closely related to the *ICC*, they argued that often it is more appropriate to use a prior distribution on the *ICC*, and information for prior distributions for the *ICC* is now becoming available (Gulliford et al. 2005). Turner et al. (2001) experimented with different prior distributions and showed that the estimate of the treatment effect is not entirely robust to the distributional assumptions of the model and suggested caution in using the conventional normality assumption. They showed that the variance components tend to be underestimated when using the non-Bayesian approach. Thompson et al. (2004) and Clark and Bachmann (2009) used Bayesian methods to analyze binary outcomes. They looked at two aspects. Firstly, they looked at rate ratios and rate differences. The latter are particularly important for economic analyses. They showed that use of Bayesian methods facilitated looking at differences in rates. Secondly they looked at the effect of different prior distributions on the outcome. Both sets of authors found that the choice of a prior distribution could have a significant effect on the treatment estimate.

10.3.5 Modeling in Matched Pair Designs

Thompson et al. (1997) replaced standard modeling approaches by techniques borrowed from meta-analysis. Thus an intervention/control pair replaces an individual clinical trial of a meta-analysis. This is essentially equivalent to relying on between-stratum information to estimate ρ under the assumption of no intervention by stratum interaction. An attractive feature of this is that the forest plot can show which pairs appear to be outliers. However, this approach requires a large number of strata (pairs) to ensure its validity and is therefore not applicable to many community intervention studies. The meta-analysis method as applied to the matched pairs design was extended to binary data by Alexander and Emerson (2005) using a Bayesian approach.

The strict lack of applicability of the t-test to binary outcomes in a matched pair design has led some investigators to alternatively recommend non-parametric approaches, such as Fisher's one sample randomization (permutation) test. Simulations performed by Gail et al. (1996) showed that inferences for matched pair binary data using permutation procedures will have significance levels near nominal under conditions likely to arise in community intervention trials. Essentially the same conclusions were reached by Brookmeyer and Chen (1998) for person-time data arising from matched pair trials. It is useful to note, however, that the one sample permutation test requires a minimum of six pairs to yield a two-sided p-value of less than 0.05, reflecting its relatively weak power. Donner and Donald (1987) showed that a weighted paired t-test based on a logistic transformation of the crude event rates tends to be more powerful than both the permutation test and the standard paired t-test in trials having a small number of strata.

10.3.6 Advice on Methods of Analysis

A method of analysis we have not discussed here is the so-called "fixed" effect method. This involves fitting dummy variables for each of the clusters. This would be relevant if we were particularly interested in the results for particular clusters. However, in general, the clusters are just a source of variation; if the trial were run again, different clusters would be used. Thus treating clusters as fixed incorrectly removes a source of variability, and so the standard errors from this approach will be incorrect.

Heo and Leon (2005) compared different methods of analysis using Eq. 10.12: (1) full likelihood, (2) penalized quasi-likelihood, (3) generalized estimating equations and (4) fixed effects logistic regression. The third method is a marginal method which, following the discussion in Sect. 10.3.1, estimates a different population parameter than the regression coefficient in Eq. 10.10. However, it does not require one to assume a normal distribution for the Z_i. The last method does not take the ICC into account and is an invalid method in general for cluster randomized trials. However, if $\sigma_Z^2 = 0$, it may be expected to yield valid tests and efficient estimates.

Heo and Leon (2005) found the full likelihood method and the penalized likelihood methods to be similar and no worse than the fixed effects method even when the within-cluster correlations are zero. As expected, the GEE method gave biased estimates of β_1, the cluster specific parameter for the treatment effect from Eq. 10.6. They did not investigate the effects of estimating β_1^*, the population averaged parameter from Eq. 10.7. Preisser et al. (2003) showed how to apply a *PA* model to a pretest posttest cross-sectional design, where the assumption of an exchangeable correlation matrix breaks down. They stated that the GEE approach is asymptotically equivalent to the summary measure approach and quoted Mancl and DeRouen (2001) to the effect that using bias-corrected variances can yield valid test sizes even with unequal cluster sizes and with as few as ten clusters.

Ukoumunne et al. (2008) carried out a number of simulations contrasting a cluster level t-test with GEE methods, when the outcome is either the difference in proportions, the risk ratio, or the odds ratio. They found that GEE had little bias in any scale, when the number of clusters per arm was at least ten. In contrast the cluster level t-test only performed reasonably for the difference in proportions.

There are other methods for the analysis of cluster trials such as the use of the bootstrap (Carpenter and Bithell 2002) and methods which fit models in stages. Feng et al. (1996) conducted a comparison of maximum likelihood assuming a normal mixed model, GEE, a bootstrap, and a "4-stage method." The bootstrap used by the authors draws a random sample of size k from the original k clusters. Then one can use ordinary least squares to estimate the βs and repeat a large number of times. The four-stage method is non-iterative, where the first step is to estimate the βs by ordinary least squares and obtain the residuals $e_i = Y_i - X_i\hat{\beta}$. Then the e_is are regressed against the Z_is, leading to estimates of σ^2 and σ_Z^2. One can then use these estimates in a weighted least squares regression of Y_i versus X_i. For small numbers of clusters (<10 per arm) and for nearly balanced data, the bootstrap has been shown to do well, especially if one does not wish to assume normality. For larger numbers of clusters, the maximum likelihood method performed better than GEE.

In practice there is a choice of four main types of analysis: use of summary statistics, generalized estimating equations (GEE), random effects models, and Bayesian random effects models. These are summarized in Fig. 10.3.

As stated earlier, within the GEE method, there is a choice of independent error structure or exchangeable error structure. Within the random effects method, there are a number of ways of fitting models which usually give similar results. For community intervention trials with few large clusters, there is much to recommend a summary measure approach: easy to implement and to understand. If the design includes matched pairs of clusters, then if the number of pairs is less than 10, an analysis ignoring matching is likely to be worthwhile. For "small cluster size" trials with more than 20 clusters per intervention group, the GEE methods using an exchangeable correlation structure are simple and robust. With fewer clusters, one could adopt a random effects model or use a cluster level method. As with any statistical analysis, with few data there is a compromise to be made between the number of assumptions about the data and the power to test hypotheses. With random effects models, it is important to test the assumptions regarding the distribution of the random effects. As stated earlier, this is most easily done using a Bayesian approach, but this requires a degree of expertise and is less commonly done.

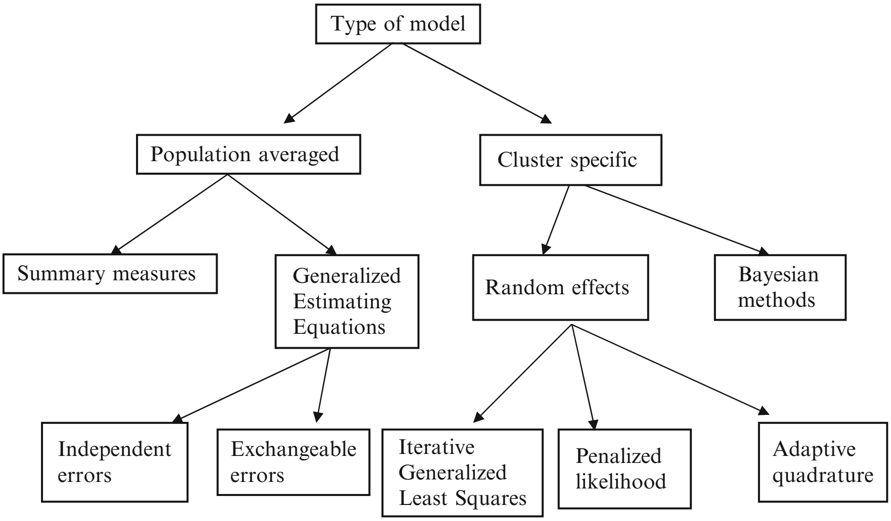


Fig. 10.3 Choices of model and fitting methods

10.4 Other Considerations

10.4.1 CONSORT Statement for Presenting Results from Cluster Randomized Trials

Cluster randomized trials are often poorly reported (Eldridge et al. 2004). The original CONSORT statement was an attempt to improve reporting of individually randomized trials. This statement was subsequently adapted for cluster randomized trials and revised by Campbell et al. (2004). The most important distinctions from the original CONSORT statement are (1) to give a rationale for adopting a cluster design, (2) to describe how the effects of clustering were incorporated into the sample size calculations, (3) to describe how the effects of clustering were incorporated into the analysis, and (4) to describe the flow of both clusters and participants through the trial, from assignment to analysis. Thus, in a primary care trial, for example, one would like to know how any primary care groups were approached, how many agreed to participate, how many were randomized, and how many dropped out during the trial, as well as the characteristics of the patients in the study. The most up-to-date statement is available at www.consort-statement.org.

10.4.2 Clustering by Therapist

Investigators may need to contend with clustering of subjects' responses even for trials using individual randomization. For example, patients may be randomized as they enter a trial by whether or not they will receive a new intervention (say acupuncture). However, there may be a limited number of acupuncturists and so a single acupuncturist may treat a number of patients. The outcomes will thus be affected not only by whether the patient received acupuncture but by which acupuncturist treats them. The model is

$$Y_{ijk} = \beta_0 + \beta_1 X_i + \gamma Z_k + \varepsilon_{ij_,}$$

where the subscript k indicates the effect of different acupuncturists and $Z_k = 0$ for the control group. However, it is important to note that the subjects in the control group are not naturally clustered. Issues in the analysis of these trials have been covered by Roberts and Roberts (2002). Lee and Thompson (2005) discussed a Bayesian approach to the analysis of such trials and pointed out how taking account clustering in the analysis can affect the results.

10.4.3 Compliance and Recruitment

We now discuss some potential sources of bias that are peculiar to cluster randomization trials. As we discussed in Sect. 10.2.2.1, if the subjects in a trial are newly diagnosed patients and the intervention is some new approach to treating a disease, then it is possible that practitioners in the intervention arm, being newly educated about this disease, may be more likely to diagnose the disease (Campbell 2000). This may lead to serious problems of selection bias if patients in the intervention arm have less serious disease than patients in the control arm. Trials should be analyzed by what is known as the "intention to treat" (ITT) principle. This means that patients randomized to a particular treatment will be analyzed as if they received that treatment, irrespective of their actual treatment. This is a so-called "pragmatic" approach which attempts to reflect what will happen in practice, when patients will not necessarily comply with treatment. An ITT approach is appropriate when compliance varies over clusters but varying compliance has major implications for any attempt at casual modeling. Loeys et al. (2001) demonstrated how to use standard GEE and random effects models to allow for variable compliance among clusters.

From experience, factors that improve compliance include building a "team spirit" within a cluster, regularly communicating to the patients about the trial, and providing the control group access to the intervention after the trial. For example, in the Hampshire Depression Trial (Thompson et al. 2000), general practitioners (GPs) in the intervention were trained to recognize depression in their patients. GPs in the control group were offered training when the trial was over, and this increased their willingness to stay in the trial.

10.4.4 Software

The selection of the general packages which can fit these models discussed in this chapter is given in Table 10.4. Stata (Statacorp 2009) has simple commands

Table 10.4 Web addresses for software packages

Name	URL
MLwiN	http://www.bristol.ac.uk/cmm/software/mlwin/
R	http://www.r-project.org/
SAS	http://www.sas.com/technologies/analytics/statistics/
Stata	http://www.stata.com/
WinBUGS	http://www.mrc-bsu.cam.ac.uk/bugs/

for applying a cluster robust standard error and for checking the distribution of the residuals. These can also be accomplished in R, but it is not as easy to use. MLwiN enables more than two-level clustering (e.g., by pupil by class by school) or therapist by patient by time within patient. It also can fit models using Markov Chain Monte Carlo methods which enable a Bayesian approach. WinBUGS can be used to analyze trials using Bayesian methodology with prior distributions for the parameters. SAS is particularly flexible for mixed models.

10.5 Conclusions

10.5.1 Review

Cluster randomized trials are an order of magnitude more complicated than ordinary randomized controlled trials. If the risk of contamination is low, then an investigator would be well advised to consider whether an individually randomized trial might be more efficient. However, the last 10 years have seen a flourishing of research into cluster randomized trials and they are now better understood and can be analyzed relatively easily using common software. There is still a need for information about likely values of the intracluster correlation coefficient for common outcomes and clusters so that trials can be planned with more precision. Trial design is still comparatively simple, and research is needed on issues such as group sequential trials where interim analysis can inform the future design of the trial.

10.5.2 Further Reading

The standard text books on cluster (or group) randomized trials are those by Murray (1998); Donner and Klar (2000) and Eldridge and Kerry (2012). A recent book by Hayes and Moulton (2009) emphasizes the use of cluster trials in infectious diseases, particularly in developing countries. There have been a number of reviews of cluster randomized trials. Klar and Donner (2001) and Donner and Klar (2004), following on from their book (Donner and Klar 2000), reviewed developments up until that time and suggested areas of further research. Murray et al. (2004) reviewed methods in public health. Methodological developments for cluster randomized trials have also been reviewed more recently by Campbell et al. (2007).

References

Alexander N, Emerson P (2005) Analysis of incidence rates in cluster-randomized trials of interventions against recurrent infections, with an application to trachoma. Stat Med 24:2637–2647

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88:125–134

Brookmeyer R, Chen Y-Q (1998) Person-time analysis of paired community intervention trials when the number of communities is small. Stat Med 17:2121–2132

Campbell MJ (2000) Cluster randomized trials in general (family) practice research. Stat Method Med Res 9:81–94

Campbell MJ (2008) Should we use relative risks or odds ratios in cluster randomized trials with binary outcomes that have high proportions? J Epidemiol Community Health 62(Suppl 1):A24

Campbell MJ, Donner A, Klar N (2007) Cluster randomized trials and *Statistics in Medicine*. Stat Med 26:2–19

Campbell MK, Elbourne DR, Altman DG (2004) CONSORT statement: extension to cluster randomized trials. BMJ 328:707–708

Carpenter J, Bithell J (2002) Bootstrap confidence intervals: when, which what? A practical guide for medical statisticians. Stat Med 19:1141–1164

Clark AB, Bachmann MO (2009) Bayesian methods for analysing cluster randomized trials with count outcome data. Stat Med 29:199–209

Davies MJ, Heller S, Skinner TC, Campbell MJ, Carey ME, Cradock S, Dallosso HM, Daly H, Doherty Y, Eaton S, Fox C, Oliver L, Rantell K, Rayman G, Khunti K (2008) Effectiveness of the diabetes education and self management for ongoing and newly diagnosed (DESMOND) programme for people with newly diagnosed type 2 diabetes: cluster randomized controlled trial. Br Med J 336:491–495

Diehr P, Martin DC, Koepsell T, Cheadle A (1995) Breaking the matches in a paired t-test for community interventions when the number of pairs is small. Stat Med 14:1491–1504

Donner A, Donald A (1987) Analysis of data arising from a stratified design with the cluster as unit of randomization. Stat Med 6:43–52

Donner A, Klar N (2000) Design and analysis of cluster randomization trials. Arnold, London

Donner A, Klar N (2004) Pitfalls of and controversies in cluster randomized trials. Am J Public Health 94:416–422

Donner A, Birkett N, Buck C (1981) Randomization by cluster: sample size requirements and analysis. Am J Epidemiol 114:906–914

Donner A, Taljaard M, Klar N (2007) The merits of breaking the matches in community intervention studies: a cautionary tale. Stat Med 9:2036–2051

Efron B, Tibshirani R (1998) An introduction to the bootstrap. CRC Press, Boca Raton/New York/Abingdon

Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC (2004) Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. Clin Trials 1:80–90

Eldridge S, Kerry S, Ashby D (2006) Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. Int J Epidemiol 35:1292–1300

Eldridge S, Kerry S, Torgerson D (2010) Bias in identifying and recruiting participants in cluster randomised trials: what can be done? BMJ 340:36–39

Eldridge S, Kerry S, (2012) A practical guide to cluster randomized trials in Health Service Research. Wiley, Chichester

Feldman HA, McKinlay SM (1994) Cohort versus cross-sectional design in large field trials: precision, sample size, and a unifying model. Stat Med 13:61–78

Feng Z, Grizzle JE (1992) Correlated binomial variates: properties of estimator of intraclass correlation and its effect on sample size calculation. Stat Med 11:1607–1614

Feng Z, McLaran D, Grizzle J (1996) A comparison of statistical methods for clustered data analysis with Gaussian error. Stat Med 15:1793–1806

Feng F, Diehr P, Yasui Y, Evans B, Beresford S, Koepsell TD (1999) Explaining community level-variance in group randomized trials. Stat Med 18:539–556

Flynn TN, Whitley E, Peters TJ (2002) Recruitment strategies in a cluster randomized trial – cost implications. Stat Med 21:397–405

Freedman LS, Green SB, Byar DP (1990) Assessing the gain in efficiency due to matching in a community intervention study. Stat Med 9:943–953

Friede T, Kieser M (2006) Sample size recalculation in internal pilot study designs: a review. Biom J 48:537–555

Gail MH, Byar DP, Pechacek TF, Corle DK (1992) Aspects of statistical design for the community intervention trial for smoking cessation (COMMIT). Control Clin Trials 13:6–21

Gail MH, Mark SD, Carroll RJ, Green SB, Pee D (1996) On design considerations and randomization-based inference for community intervention trials. Stat Med 15:1069–1092

Goldstein H (2002) Multilevel statistical models, 3rd edn. Wiley, Chichester

Grosskurth H, Mosha F, Todd J, Mwijarubi E, Klokke A, Senkoro K, Mayaud P, Changalucha J, Nicoll A, Ka-Gina G (1995) Impact of improved treatment of sexually transmitted diseases on HIV infection in rural Tanzania: randomized controlled trial. Lancet 346:530–536

Gulliford MC, Adams G, Ukoumunne OC, Latinovic R, Chinn S, Campbell MJ (2005) Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. J Clin Epidemiol 58:246–251

Hayes RJ, Moulton LH (2009) Cluster randomized trials. CRC press, Boca Raton/New York/Abingdon

Heo M, Leon AC (2005) Comparison of statistical methods for analysis of clustered binary observations. Stat Med 24:911–923

Hsieh FY (1988) Sample size formulae for intervention studies with the cluster as unit of randomization. Stat Med 8:1195–1201

Huber PJ, Ronchetti EM (2009) Robust statistics, 2nd edn. Wiley, Hoboken

Kerry SM, Bland JM (2001) Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. Stat Med 20:377–390

Kinmonth AL, Woodcock A, Griffin S, Spiegal N, Campbell MJ (1998) Randomized controlled trial of patient centred care of diabetes in general practice: impact on current well being and future disease risk. Br Med J 317:1202–1208

Klar N, Darlington G (2004) Methods for modelling change in cluster randomization trials. Stat Med 23:2341–2357

Klar N, Donner A (1997) The merits of matching in community intervention trials. Stat Med 16:1753–1764

Klar N, Donner A (2001) Current and future challenges in the design and analysis of cluster randomization trials. Stat Med 20:3729–3740

Lake S, Kamman E, Klar N, Betensky R (2002) Sample size re-estimation in cluster randomization trials. Stat Med 21:1337–1350

Lee KJ, Thompson SG (2005) The use of random effects models to allow for clustering in individually randomized trials. Clin Trials 2:163–173

Lewsey JD (2004) Comparing completely and stratified randomization designs in cluster randomized trials when the stratifying factor is cluster size: a simulation study. Stat Med 23:897–905

Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22

Loeys T, Vansteelandt S, Goetghbeur E (2001) Accounting for correlation and compliance in cluster randomized trials. Stat Med 20:3753–3767

Machin D, Campbell MJ, Tan S-B, Tan S-H (2008) Statistical tables for the design of clinical studies. Wiley-Blackwell, Chichester

Mancl L, DeRouen TA (2001) A covariance estimator for GEE with improved small sample properties. Biometrics 57:126–134

Martin DC, Diehr P, Perrin EB, Koepsell TD (1993) The effect of matching on the power of randomized community intervention studies. Stat Med 12:329–338

Matthews JNS, Altman DG, Campbell MJ, Royston P (1990) Analysis of serial data using summary measures. Br Med J 300:230–235

Moerbeek M (2006) Power and money in cluster randomized trials: when is it worth measuring a covariate? Stat Med 25:2607–2617

Mills S, Campbell MJ, Waters WE (1986) Public knowledge of AIDS and the DHSS advertisement campaign. Br Med J 293:1089–1090

Murray DM (1998) The design and analysis of group-randomized trials. Oxford University Press, Oxford

Murray DM, Vernell SP, Blitstein JL (2004) Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health 94:423–432

Neuhaus JM, Jewell NP (1993) A geometric approach to assess bias due to omitted covariates in generalized liner models. Biometrika 80:807–815

Nixon RM, Thompson SG (2003) Baseline adjustments for binary data in repeated cross-sectional cluster randomized trials. Stat Med 22:2673–2692

Pan W, Wall MM (2002) Small sample adjustments in using the sandwich variance estimator in generalized estimating equations. Stat Med 21:1429–1441

Preisser JJ, Young ML, Zaccaro DJ, Wolfson M (2003) An integrated population average approach to the design, analysis and sample size determination of cluster unit trials. Stat Med 22:1235–1254

Roberts C, Roberts SA (2002) Design and analysis of clinical trials with clustering effects due to treatment. Clin Trials 2:152–162

Soncini JA, Maserejian N, Trachtenberg F, Tavares M, Hayes C (2007) The longevity of amalgam versus compomer/composite restorations in posterior primary and permanent teeth. J Am Dent Assoc 138:763–772

Spiegelhalter DJ (2001) Bayesian methods for cluster randomized trials with continuous response. Stat Med 20:435–452

Statacorp (2009) Stata release 11. Statistical software. Statacorp LP, College Station

Thompson C, Kinmonth AL, Stevens L, Peveler R, Stevens R, Ostler K, Pickering RM, Baker NG, Henson A, Preece J, Cooper D, Campbell MJ (2000) Effects of a clinical practice guideline and practice based education on the detection and outcome of depression in primary care: Hampshire depression project randomised controlled trial. Lancet 355:185–191

Thompson SG, Pyke S, Hardy RJ (1997) The design and analysis of paired cluster randomized trials: an application of meta-analysis techniques. Stat Med 16:2063–2980

Thompson SG, Warn DE, Turner RM (2004) Bayesian methods for analysis of binary outcome data in cluster randomized trials on the absolute risk scale. Stat Med 23:389–410

Turner RM, Omar RZ, Thompson SG (2001) Bayesian methods of analysis for cluster randomized trials with binary outcome data. Stat Med 20:453–472

Turner RM, Prevost AT, Thompson SG (2004) Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. Stat Med 23:1195–1214

Ukoumunne OC, Thompson SG (2001) Analysis of cluster randomized trials with repeated cross-sectional binary measurements. Stat Med 20:417–433

Ukoumunne OC, Carlin JB, Gulliford MC (2007) A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials. Stat Med 26:3415–3428

Ukoumunne OC, Forbes AB, Carlin JB, Gulliford MC (2008) Comparison of the risk difference, risk ratio and odds ratio scales for quantifying the unadjusted intervention effect in cluster randomized trials. Stat Med 27:5143–5155

Wight DG, Raab GM, Henderson M, Abraham C, Buston K, Hart G, Scott S (2002) Limits of teacher delivered sex education: interim behavioural outcomes from randomized trial. Br Med J 324:1430–1433

Community-Based Health Promotion 11

John W. Farquhar

Contents

J.W. Farquhar
Stanford Prevention Research Center, Stanford University School of Medicine, Stanford,
CA, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_33,

11.1 Introduction

This chapter describes theories and methods underlying successful *comprehensive community-based health promotion* and provides five examples. These studies represent well the field of *experimental* epidemiology, involving the total populations of communities, and often done following insights derived from *observational* epidemiology or from *randomized clinical trials* in smaller populations. Three of these were designed to reduce risk factors for cardiovascular disease (CVD) and relied heavily on locally available channels of mass communication in addition to community organizing and other education methods, relatively low-cost approaches with the potential to reach and change lifestyle behaviors of entire populations, in contrast to traditional individual or group counseling. One example, studying alcohol-involved trauma, used only community organizing to promote adherence to existing laws, rather than public education for behavior change. A second example, designed to decrease consumption of dietary trans fats, used primarily new laws and regulations. By community organizing, we mean the process of enlisting community leaders in support of project goals, and also in insuring their continued support. The Stanford Prevention Research Center (SPRC), beginning in 1972, pioneered development of the intervention methods of comprehensive community-based CVD prevention and other methods of health promotion and chronic disease prevention. SPRC was connected to four of the five examples, either as initiator (for two studies) or collaborator (for two studies). This review also describes the theoretical background, methods of intervention, the past history of such studies, the cultural basis for barriers to change, and lessons learned for the future. Although the examples given are derived from studies in high-income countries, the underlying theories and the intervention principles fully apply to countries at all development levels (Institute of Medicine (IOM) 2010).

A community-based program is defined as one organized locally and promoted through the community's institutions and communication channels. The traditional definition of a "community" is used in this review (a residential area with legally defined geographical boundaries, where a *local* governmental system regulates many aspects of schools, businesses, transportation, law enforcement, and recreational activities). A community is ordinarily last in a nation's regulatory chain, where education must ultimately occur, although for rural areas (in the United States) the county becomes the governing agent for education.

This chapter will address the following issues:

1. The advantages are presented for community-wide interventions rather than for more limited locales, such as clinics, hospitals, worksites, or schools.
2. The theories underlying successful community-based projects are described, including:
   a. Community organizing theory and its relationship to community self-development and diffusion of innovation theories

   b. The health communication-behavior change theory and its relationship to social cognitive theory, social marketing, and to other determinants of successful use of mass media for health promotion

3. The methods needed for success are presented, including:
   a. Message design through formative research
   b. Process analysis for comprehension of causes for change
   c. The role of community activism, advocacy, laws, and regulations

4. The history of four recent decades of comprehensive community-based health promotion for cardiovascular disease prevention is described.
5. The cultural basis for barriers to success is outlined.
6. Lastly, a *Master Plan* for achieving success in this type of health promotion is presented.

11.2 Advantages of a Total Community Approach

Health promotion in schools, worksites, and clinics has a long history of individual success, but synergistic interactive effects can occur when they are imbedded in a *total community* campaign that adds inherently cost-effective mass media and environmental change (Schooler et al. 1997). Evidence for synergism was given by Rogers (1983), who found that diffusion of innovations within a community accelerates when adoption of the innovation reaches about 20% of the population – thus only comprehensive total community education programs have the capacity to achieve such effects. Lifestyles, such as tobacco use, dietary habits, and exercise patterns, so strongly influenced by custom and by the media in developed countries, cannot be countered through simple means. Community-wide approaches fit the public health model because the usual medical model can neither prevent most chronic disease nor reach the entire population in need. By the "medical model," we mean the aspects of a nation's health care system, focused on the individual, that rely primarily on clinic outpatient and hospital services provided by physicians. The total community model provides influence through *locally produced* electronic and print mass media and through worksites, recreation sites, libraries, schools, medical, hospital, and pharmacy settings, and social gatherings of many sorts. It offers opportunities for health-promoting regulations, such as providing opportunities for physical activity for all, school fitness classes, healthful school lunches, "point of purchase" nutrition information in local food markets, alcohol sales limits, and preventing tobacco product marketing to children.

Studies at the SPRC have shown that only *multiple and persistent* influences produce meaningful changes in the dietary, exercise, or tobacco-use behavior of adults, adolescents, and children. For adolescents, always resistant to health behavior change, the following influenced success: parent and teachers whose personal habits allow them to be supportive role models; amount and quality of school-based education on tobacco, fitness, and nutrition; peer influence; and amount, duration, and quality of community-wide health education. Also, certain personal characteristics were influential – for example, the presence before the education's onset of *self-efficacy* toward one's behavior-change abilities (Fortmann et al. 1995).

11.3 Underlying Theories

Community-based health education carried out by SPRC and by some analogous projects has been guided by two major theories (community organizing and health communication-behavior change), with success dependent upon a judicious blending of the two (Flora et al. 1989; Farquhar et al. 1991).

11.3.1 Community Organizing Theory

Community organizing theory describes methods of identifying the health problem (and resources needed or available), mobilizing the community's opinion leaders and organizations, gaining populace support, forming coalitions, launching and maintaining education programs, achieving regulatory changes, and *empowering communities* to reach and maintain their goals. Community organizing for health requires continued attention similar to the accepted role of political leaders – to assess periodically the needs of both governmental and non-governmental organizations, and, in the case of health promotion, to aid in planning and coordinating health promotion campaigns. Community organizing as defined here has analogies to "community self-development" as described by Green and Kreuter (1991), and it also relies on elements of diffusion theory (Rogers 1983) – which has shown how innovations are adopted through natural social networks, aided by a community's opinion leaders. Rogers' account of "failure of water-boiling in a Peruvian village" provided an excellent example of the need to identify and work with a community's opinion leaders as a prerequisite for the success of a health innovation initiated solely by self-appointed "experts" who are not seen as trustworthy by the community's residents.

11.3.2 The Health Communication-Behavior Change Theory

The health communication-behavior change theory provides the basis for designing, sequencing. and distributing messages for the total population and its subgroups based on their health needs, cultural attributes, social networks, media habits, attitudes, motivation, knowledge, and self-management skills. It describes theories underlying educational *content*, such as social cognitive theory (Bandura 1986), which is primarily based on an individual's capacity for self-directed change. Bandura's research confirms that "learning by doing" is more effective than "learning from observing" (modeling a behavior), and both are more effective than an "information-only" approach that changes knowledge alone. These principles are contained in his *social cognitive theory*, which posits that guided practice in a new behavior can lead to increased self-efficacy and to greater behavior change (Bandura 1986). Thus, "knowledge-only" campaigns have been found less effective than those that apply Bandura's recommendations.

The health communication-behavior change theory also incorporates methods of reaching the total population using principles described as "social marketing," (Kotler 1975; Lefebvre and Flora 1988). These marketing principles described first by Kotler as "product, price, and promotion" lead to insuring the relevance of the "product" (health messages) and to the low material and psychological cost of attending to the health message. Additionally, successful social marketing occurs only when the health messages are promoted and distributed efficiently to a large proportion of the populace.

Another behavior change method included within the communication-behavior change theory is the method of teaching counterarguing skills, as described by Roberts and Maccoby (1973). This method, also called "inoculation," can be presented either through the media or face-to-face and teaches the learner how to best argue against a deleterious message, such as an advertisement promoting cigarette use. This "inoculation" method was found to be very effective in prevention of adolescent substance abuse, notably for prevention of initiation of tobacco use when used in a manner that can be incorporated into comprehensive community-based health promotion (Robinson et al. 1987).

Lastly, Cartwright's (1949) pioneering work on mass persuasion principles is also quite germane to the health communication-behavior change theory. He described the need to change not only a person's knowledge and motivation, but that changes in "action structure" were also needed. These principles, when combined with Bandura's methods for self-management skills training and with certain elements of Rogers' diffusion theory, lead to a logical sequence of steps recommended for both the delivery and behavioral objectives, as derived from the health communication-behavior change theory (Table 11.1).

Table 11.1 The health communication-behavior change components (Adapted from Farquhar et al. 1991)

Communication inputs	Communication functions (for the sender)	Behavior objectives (for the receiver)
Face-to-face messages	Determine receiver's needs	Become aware
Mediated messages	Gain attention (set the agenda)	Increase knowledge
Community events	Provide information	Increase motivation and interest
Environmental cues	Provide incentives	Take action, assess outcomes
	Provide training	Maintain action, practice self-management skills
	Provide cues to action, including environmental change	Become an opinion leader (exert peer group influence)

11.4 Methods

11.4.1 Initial Steps

Problem Identification The initiating agency, group, or person can emanate either from within or outside of the community. The first task is to identify the problem. Local (community) interventions should ideally coexist with and interact with national programs and must be linked to scientific support from the national or international level, which has indicated that both CVD prevention (WHO 1986) and alcohol sales control (Holder and Wallack 1986) are problems requiring *local* action. Other than the initial steps of problem identification and formation of a coalition, most of the methods to be described are relevant only to the three CVD prevention examples described in this chapter. However, any chronic disease that requires widespread education designed to change "lifestyle" behaviors (such as exercise, diet, and cigarette use) will require analogous methods.

Coalition Formation Key political and opinion leaders as well as relevant organizations must form a coalition. This coalition must create a resource inventory, obtain populace support, and plan the intervention. Relevant organizations, particularly in developed countries, are listed below under Sect. 11.4.3 and include the following: the County Medical Society, or its equivalent, municipal hospital community affairs departments, city parks and recreation departments and voluntary health agencies (i.e., the local branches of national heart, lung, cancer and diabetes organizations). Any organization that will act as a conduit (or "channel") for the distribution of mediated instruction or classes must also be represented (i.e., television and radio stations, local newspapers, schools, churches, libraries, pharmacies, clinics, and physician and dentist offices). As the coalition grows in size, a "steering committee," preferably restricted to six members, needs to be formed to carry out more detailed planning, including formation of expert groups in education and evaluation and "task forces" assigned to particular topics (Flora et al. 1989; Farquhar et al. 1991).

11.4.2 Planning

Formative Evaluation As the term "formative" indicates, these activities form (create) the education effort and can be divided into categories of audience needs analysis, message design, pretesting of education programs, and evaluation of education programs. The needs analysis, message design, and pretesting phases are part of the social marketing aspect of health communication-behavior change theory. Also, the evaluation of education programs, although properly labeled as "formative," is also often termed "process" evaluation since it determines the contributions to success of different components of each education program (Farquhar et al. 1991). Process analysis is also done at the completion of the total program as part of "summative evaluation," although the distinction between formative and summative is sometimes quite arbitrary, depending on the intent of the evaluator (Farquhar et al. 1991).

11.4.3 Implementation

Need for a Comprehensive Approach Interventions must go beyond attempting to change knowledge, the usual goal of an educational system, by providing training in behavior-change *skills.* They must also go well beyond the individual, enlisting multiple community organizations in campaigns for change and seeking changes in the social environment and in regulations that promote access to the facilities and resources needed for healthful practices. Multiple communication channels (such as radio, television, newspapers, mass-distributed print products, schools, and the Internet) are needed to reach different subgroups, recognizing differing media usages, knowledge, and desire for change. A comprehensive intervention should involve schools, worksites, senior citizens' centers, voluntary health agencies (such as the local branches of any national organization that deals with cardiovascular disease, diabetes, and cancer), churches, and facilities for sport, recreation, and health. These organizations and others can serve as education conduits, with the community's electronic and print mass media organizations assisting in message design, content, and delivery. The Internet provides a new channel, whose community education role is now becoming better defined (Baker et al. 2003). Interactive computer learning, now becoming common in classrooms, can be designed for large groups – an emerging variant of mass media. Newer communication technologies, such as handheld computers, provide new points of entry for health promotion (King et al. 2008).

Comprehensiveness requires variety. For example, in tobacco control programs of SPRC's community campaigns, local medical clinics, dental offices, pharmacies, and libraries distributed a low-cost skills training "Quit Kit"; a local smoking cessation class was shown on television; many newspaper articles and columns appeared; local businesses supported costs of a smoking cessation contest; and all newspapers and electronic media "cross-advertised" activities designed for mass audiences (Altman et al. 1987). Formative evaluation must be continued throughout the entire implementation period of any planned intervention campaign. Success requires a well-designed mix and sequence of programs delivered through varied channels. This integration, with goals set in advance and goal changes based on early results, is analogous to a commercial marketing campaign, hence the term "social marketing." The distinction from commercial marketing is that social marketing uses marketing methods for social betterment without a commercial or profit-making intent. As described above, social marketing is the explanatory term for the needs analysis, message design, and pretesting phases of formative evaluation. These phases require message tailoring to fit any subgroup's needs and preferences, respecting cultural differences, learning styles, and preferred learning sites (Slater and Flora 1991). A message sequence should increase awareness, then, increase knowledge, and last, increase motivation and provide training in the skills needed for adoption and maintenance of a new behavior (Bandura 1986). Electronic media can carry out the first two parts of this sequence and stimulate use of the more information-dense print media of newspapers and booklets, which are inherently more effective in *skills training* than are electronic media (Flora et al. 1997).

Message Characteristics Messages must be clear, focused, and salient. Salience requires broad reaching media, arousing interest and awareness – topics must break through passive indifference engendered by the information overload of many societies and become "on the public agenda." Given the large advertising budgets of today's commercial mass media, health agencies' messages must be of sufficient production quality to compete for the public's attention. The competition for the public's attention is great indeed since the average adult in the United States in the past decade was exposed annually to 35,000 television advertisements, equaling 292 h of exposure (Fortmann et al. 1995). Formative evaluation must continue during a campaign and alter message content in accordance to the state of readiness of the population. Thus, early in a campaign, messages should increase knowledge and awareness, followed by those that increase motivation and provide skills training – with resultant behavior change (Schooler et al. 1997).

11.4.4 The Amount of Intervention Needed

The amount of intervention (the "dose" of intervention) needed depends on many factors: lesser amounts are needed in smaller communities, at earlier stages in a country's adoption of a "health innovation," and when the advocated behavior change is reasonably simple (such as mammography, hypertension screening, and immunization campaigns). Clearly, more complex changes are needed in individuals and in society's norms to alter eating or exercise patterns or to control tobacco use. Complexity in respect to nutrition arises from many sources, including long-standing cultural beliefs and practices; entrenched methods in agriculture, food production, and retailing; advertising of "unhealthful" foods; and the advent of widespread fast-food chains that are dominated by commercial interests unresponsive to local demands and needs. Few projects have measured intervention dose, except in very general terms that do not allow accurate estimates of the amount of exposure. One excellent method records the total number and duration of messages distributed over a defined time period, albeit with a defect due to lack of message quality measures (Farquhar et al. 1990). This method's use increased evaluation costs (which comprised almost two-thirds of total expenses) in the Stanford Five-City Project (FCP – see below). This expense may explain why others have not used it nearly to the same degree of completeness. However, it is clear that public health education would be served if more campaigns were analyzed this thoroughly. The following describes this method's use in the Stanford Five-City Project (see Sect. 11.5.3). The number of adults aged 18–74 were known in the communities receiving the education programs, and they were used for the analysis. The number and duration of messages from TV/radio, newspaper, other print messages, and face-to-face messages (largely classes) were enumerated each year for 5 years. The number of messages from newspaper, other print and face-to-face encounters were obtained from the education group's records. The TV broadcast hours were obtained from the Neilsen (A.C. Neilsen Co, Chicago IL) monitoring system, which records the proportion of households viewing a particular TV station for all hours for each major community area in the United States. Radio hours were obtained from the local radio station survey data. Calculations are then made for the number of messages for each education channel and their duration for the average adult in the community (Flora et al. 1989; Farquhar et al. 1990).

11.4.5 Evaluation

Comprehensive community-based campaigns aiming for widespread behavior change in the community's population need the following types of evaluation: (1) formative evaluation to plan and test messages (as described in Sect. 11.4.2), (2) process evaluation to study the effects of each individual educational activity, and (3) summative evaluation to determine the effects of the campaign at different levels (the individual, in organizations, and in community's social or physical environment). Formative evaluation determines the likelihood that a message, class, or group activity will reach the intended audience, and requires measures of audience segmentation, trials of effects of message design, evaluation of form and content of mediated message delivery, and other features of social marketing (see Sect. 11.4.3) (Lefebvre and Flora 1988; Slater and Flora 1991; Cirksena and Flora 1995; Williams and Flora 1995). Process evaluation examines the success or failure of a program component and also allows insight into why it succeeded (or failed). Process analysis requires a dissection of relative effectiveness of multiple channels, differentiating their reach, specificity, and impact (Schooler et al. 1998). Process analysis allows insight into "how" and "why" a change campaign worked, or failed – finding the reasons explaining outcomes measured by summative evaluation (Flora 1991). As mentioned in Sect. 11.4.2, process analysis may be classified under either "formative" or "summative," depending on timing and purpose (Flora et al. 1989; Farquhar et al. 1991). Summative evaluation methods are generally more rigorous than either formative or process evaluation methods and, in the case of chronic disease prevention interventions, often will entail measures of physiological states (such as blood pressure), as well as attitudes, knowledge, or behavior (such as cigarette use) (Farquhar et al. 1977, 1990). The unit of analysis is commonly the individual, although both the Stanford Three Community Study (Williams et al. 1981) and the Five-City Project (Farquhar et al. 1990) also analyzed risk factor change by the more conservative method, using longitudinal regression analyses with the community as the unit of analysis (recognizing time-by-town variance). This is an example of multilevel analysis, or as a nested model, with the survey data on smoking rates, blood pressure and other variables representing the individual level, and the mean levels of these variables for the entire population representing the community level. When, as was the case in both the TCS and FCP, the two levels have congruent outcomes, causal inference is strengthened (Farquhar et al. 1977, 1990). More complex multilevel analyses are needed for projects with more complex designs. For example, a multi-school project designed to decrease student's tobacco use was analyzed at three levels (individual, classroom, and school) by a random regression model which was shown to be better able to discern variance missed by ordinary regression techniques (Hedeker et al. 1994). In future intervention studies, especially when community sub-units such as schools or worksites are being studied, multilevel methods should be used with greater frequency, and thus allow greater insight into both causality and process of change (Hedeker et al. 1994).

11.4.6 Additional Methods Needed

Successful community-based health promotion requires effective leaders, community activists with the courage and charisma to advocate health innovations. Advocacy campaigns derived from international, national, state, or provincial sources can provide a local activist leader with the popular support to fight entrenched bureaucrats who defend the status quo. Tobacco control in Australia and the United States provides examples. National and state advocacy groups with access to mass media created a strong mass movement for change, allowing advocates to enlist popular support for local tobacco control measures. In both California and Australia's State of Victoria, this popular support led to statewide increases in tobacco taxes, with some retained for education against tobacco, a measure that had been resisted by state legislators who had long been influenced by tobacco lobbyists – an example of community activism moving up the "ladder" of bureaucracy to a higher political level (Victorian Health Promotion Foundation 1994; Pierce et al. 1994; The Catalonia Declaration 1996).

Changes in policies, laws, and regulations (PLR) are needed for success, especially for long-term success. Local PLR can affect alcohol and tobacco sales and create environments that improve nutrition and enhance physical activity. However, national, state, or provincial actions can magnify local PLR and education efforts on topics such as tobacco taxation, automobile seat-belt laws, food and drug safety, school nutrition, school physical activity policies, and (in the United States) laws on firearms. As described above, widespread popular attitude changes in numerous communities can also affect the political process at the state, province, or federal level. Tobacco control is a case in point, with major decreases in tobacco use in many countries in the past four decades (NHS Information Service 2007; Fiore and Jaen 2008). These successes, and their accompanying political ferment, were largely responsible for the passage of the WHO Framework Convention on Tobacco Control in 2003, with ratification by more than 100 countries by 2007 (Koh et al. 2007).

11.5 History of Comprehensive Community Health Promotion

The history of the past four decades described in this chapter is devoted largely to eight formal research projects designed to affect cardiovascular disease (CVD) risk factors, and one designed to limit alcohol use in a community. All but one involved entire populations of at least one *education* community, compared to at least one *control* community. One of those was restricted to a before/after analysis in a single large community (New York City). Five projects are described in greater detail to provide examples of the methods and results. Dissemination worldwide into practical applications of community organizing and mass communication technologies, derived in part from these research projects, occurred throughout these four decades, with a relatively small number using research methods analogous to those described here.

11.5.1 The First Two Examples: The First Decade

The first and second example, the Stanford Three Community Study (TCS), in three small agricultural marketing towns in California (total population 45,000), and the North Karelia Study (NKS), in two adjoining predominately rural Finnish counties (North Karelia population about 180,000), each began in 1972.

TCS, the first Stanford project, was carried out from 1972 to 1975 in both English and Spanish, comparing effects of mass media alone in one community and mass media plus ten-session risk reduction classes for some high-risk adults in a second, with a third as a control (Farquhar et al. 1977; Schooler et al. 1997). Groups exposed to varied education amounts showed a dose-response change in smoking, blood pressure, and blood cholesterol, with a proportionately larger effect in the Spanish-speaking residents than in the Anglo majority. This minority population outcome required intervention resources in Spanish to be proportionately larger than those provided in English to the Anglo majority. A composite CVD risk reduction of about 23% and 30% occurred in the mass media-only and mass media-plus classroom conditions, respectively. The risk score (a probability) was derived from each adult's "before-after" risk levels (age, gender, systolic blood pressure, blood total cholesterol level, cigarette use, and relative weight). These risk parameters were entered into a multiple logistic regression model to predict the 12-year future probability of a coronary heart disease event (myocardial infarction, sudden death, or angina pectoris). The scoring system was based on coronary events experienced by adults in the long-term prospective Framingham Heart Study (Truett et al. 1967). Thus, a relatively modest amount of mass media (about 30 television and radio "spots," weekly newspaper columns on heart health, and four separate mass mailings of booklets) was sufficient to change the population's body weight, cholesterol and blood pressure levels, and smoking prevalence. The education followed the principles outlined in Sect. 11.3 for the community organizing and health communication-behavior change theories. As an example of community organizing, close cooperative relationships were created with two influential opinion leaders (a local Anglo physician and the Hispanic program director of the local Spanish language radio station). As an example of the social marketing aspect of the health communication-behavior change theory, message design for the smoking cessation print materials were different in the English and Spanish languages, as dictated by formative evaluation. Both the eight session classes held for a high-risk subset of the population and the print products mailed to households incorporated some of Bandura's self-management principles, such as building awareness, making a clear commitment (i.e., a written "contract"), and adopting gradual, stepwise changes in behavior designed to increase confidence in one's ability to achieve the desired behavior change (Bandura 1986).

NKS evaluated two matched, rural Finnish counties that contained many villages with farming and lumbering as the main occupations. North Karelia (about 180,000 population) received an education campaign that began in 1972, continuing to the present and expanded to the entire country. After 10 years, CVD risk factor changes comparable to the TCS occurred, and significant net reductions in CVD events occurred (Puska et al. 1985; Schooler et al. 1997). This study was marked by extensive community organizing, resulting in strong partnerships with residents and their organizations. The NKS influence on its country's policies was unparalleled among the CVD projects, providing its most important lesson – that a well done project led by respected scientists can move an entire country. As examples, Finland's food and agricultural industries made large changes: Fertilizers were supplemented with selenium (a substance low in Finland's soil that is needed for health), milk pricing was changed (based on protein instead of fat content), programs were created to replace dairy farms with berry farms, a new canola industry replaced jobs lost in dairying, and increased production of low-saturated fat foods occurred (Puska et al. 1985; Puska 1995; The Catalonia Declaration 1996). In 1972, population-wide nutrition change and smoking cessation interventions were an innovation internationally, which may partly explain the success of these two pioneering programs (TCS and NKS).

11.5.2 Early International Diffusion, 1977–1983

Studies similar to the TCS were done in Italy (The Martignacco Project, 1977–1983, one treatment, mass media and screening, one control – CHD risk fell in men only); Australia (The North Coast Project, 1978–1980, similar to TCS, one mass media, one mass media plus classes – effects on smoking, greater in the mass media "plus" community); Switzerland (two treatment, two control pairs, German and French speaking, mass media, classes, environment changes – effects on smoking, blood pressure, and obesity); and South Africa (similar to TCS, one mass media, one mass media plus classes plus community events – decreased CHD risk, blood pressure, smoking; both treatments were equivalent). All four studies, reviewed in Schooler et al. (1997), reported significant risk factor changes, adding evidence for the effectiveness of the TCS model, namely, that a predominately community-wide mass media campaign is effective, but that supplemental face-to-face instruction usually adds some extra effects (Farquhar et al. 1991). It is noteworthy that all of these studies were done in rather small communities (population sizes of about 12,000–15,000 residents); thus, these effects may be more difficult to achieve in larger and more complex communities.

11.5.3 The Second and Third Decades: Projects Begun in the 1980s and 1990s

The third example is the Stanford Five-City Project (FCP) (USA). Its intervention phase from 1980 to 1986 extended TCS methods to larger populations (total population of about 360,000) with multifactor CVD prevention directed at two northern California cities. There were three control cities (Farquhar et al. 1990). It differed from TCS in the larger size of the communities, in greater use of community organizing, and in greater collaboration with the communities' health, media, and education organizations in planning and implementing programs. It was similar in generous use of mass media, both print and electronic. The *proportion* of messages received by the average adult over 5 years of education was as follows: television and radio 67%, newspaper 28%, other print (such as booklets and "tip" sheets) 4%, and face-to-face 1%. Duration of exposure over 5 years was as follows: television and radio 35%, newspaper 18%, other print 41%, and face-to-face 5%. It was unique in measuring total dose of education (about 5 h/year and about 100 episodes of exposure/year to all forms of media and classroom education) (Flora et al. 1989; Farquhar et al. 1990). This means that an "average adult" who has followed the whole program on mass media and in classrooms would have experienced 100 single exposure episodes (including TV, radio, newspaper or other print, workshop training, or community sponsored exercise sessions) which add up to 5 h duration in total in each year.

FCP's initial year of television messages (defined as "high reach/low involvement") stimulated the public's use of print media, which supplied more effective training than from television in the skills needed for smoking cessation, healthful food purchasing or preparation, learning appropriate exercise habits or ways to lose or control body weight (Flora et al. 1997). However, television can be quite effective in skills training. For example, the FCP presented an 8 week "live" TV broadcast of a local smoking cessation class, which contained considerable modeling of the smoking cessation skills being learned by the class attendees (such as behavioral problem solving, self monitoring, tapering, goal setting, deep muscle relaxation, and group social support) (Altman et al. 1987). The composite antismoking campaign of a "quit smoking" contest, the television and face-to-face classes, and a widely distributed print "Quit Kit" were analyzed separately (Altman et al. 1987). Results were comparable to the TCS (about a 15% fall in Framingham composite risk of CVD) (Truett et al. 1967), with a major impact on blood pressure (4–5%) and smoking (13%) (Farquhar et al. 1990). Health bureaucracies, usually timid, should gain courage from the FCP's "David and Goliath" demonstration that only 3 h/year of high quality television health education can counteract the public's exposure to about 100 h/year of television advertising devoted to unhealthy nutrition. The exposure of the US population is estimated as 292 h of TV advertisement based on the Neilsen monitoring system (cf. Sect. 11.4.4). Michael Jacobson of the Center for Science in the Public Interest estimated that about one-third of these are for "unhealthy" nutrition (personal communication). Almost all TV nutrition advertising is "unhealthy" since the food industry is so inclined. For example, after the nutritionists and preventive medicine scientists succeeded in the 1970s, through education, in decreasing egg and butter intake, the counterattack began to bring the US population back to their previous unhealthy consumption. Although all preceding CVD studies showed effects in small towns and/or rural districts, FCP showed benefit in *cities* (with populations as high as 100,000). Also, these effects occurred despite the advent (at least in the state of California) in the 1980s of dual working families and increasing public use of fast food, factors that made achieving nutrition-behavior change more difficult.

FCP's modest resources and greater effects, as compared to similar projects, support the benefits of the mass media presentation components of the health communication-behavior change theory, including its adaptation of the skills training aspects of Bandura's social cognitive theory. Perhaps the most practical lesson to policymakers is that adult quitters saved over three times more money from their cessation of cigarette purchases (2,860 quitters @ $274/quitter/year) than the cost of the entire population-wide campaign ($4/adult/year) – savings retained by the individuals of the community, not counting savings in long-term health costs and short-term decreases in absenteeism from the individual's employment. Also, the relatively low cost of the intervention is highlighted by the fact that the total yearly cost of cigarettes to the 22,000 smokers (over $6 million) was 23 times greater than the yearly campaign costs (about $260,000/year). Lastly, the communities expanded the health promotion activities of the county's Department of Health and adopted FCP's technologies, later applying them to seat-belt promotion, changes in saturated fat use by restaurants and violence and adolescent pregnancy prevention.

Other successful CVD projects, both large and small, occurred in these decades in the United States, Sweden, Denmark, Canada, Germany, the Czech Republic, and China. Also, World Health Organization-sponsored projects began in about 23 other countries (Schooler et al. 1997) and have expanded beyond that since 1997. In all instances, they borrowed heavily from the experiences of the three exemplars and in many instances received training from either the Stanford or the North Karelia groups. In some of these projects, changes seen were less than was the case in the Stanford and North Karelia projects. Although it is difficult to know the reasons for a relative lack of success (e.g., the magnitude of the interventions was incompletely described), in some instances, the cause appeared to be related to inadequate use of the inherently cost-effective mass media.

11.5.4 Community Projects in Other Health Topics

Interventions on alcohol, mammography, tobacco control, immunization, motor vehicle injuries, and HIV/AIDS are prominent examples of other community-based projects done in developed countries, but with fewer well-controlled studies than in CVD. Also, many community studies of CVD have been done in developing countries, where effective recruiting methods and effective mass media use (often using radio messages) have been reported. A review of these is beyond the scope of this chapter, but may be found in the extensive review of measures to improve cardiovascular health in developing countries (IOM 2010). One well-controlled US study, Preventing Alcohol Trauma (PAT),[1] is the fourth example. This 5-year study of three US communities showed a 10% decrease in alcohol-related traffic injuries and a 50% decrease in adolescent alcohol use (Holder et al. 1997). The traffic injuries were analyzed for 3 years prior to the study and were compared to police and hospital records for the 5-year study duration. The alcohol use data were compiled from various records, prior to and during the study. These records included (1) arrest records of adolescents under the age of 18 for drinking while driving and (2) sales records from retail liquor stores and bars or taverns, using underage adolescents as attempted purchasers (to test the system). Coalition building and organizational behavior change among the police, alcohol sales outlets, and alcohol servers (such as bar tenders and restaurant personnel) were the main interventions. The public responded to a fear, instilled through publicity, of the penalties of greater enforcement of existing laws on underage alcohol purchases or drinking and driving. Therefore, in contrast to the three CVD exemplars, PAT showed that *major* public education is not needed for large public behavior changes. PAT estimated a cost saving of about $3 for every dollar invested, a number close to that found in many worksite health promotion studies, where the “return on investment” was $3–$6 for each dollar invested, measured in 2–5 years of the program (Aldana 2001). The cost savings came from decreased medical costs secondary to the reduced rate of alcohol-related automobile injuries. PAT found their communities had the required infrastructure for the campaigns, requiring only training provided by one indigenous community coordinator, a part-time clerk, an imaginative plan, and the will to proceed. A fifth example, from a very recent study, showed remarkable success from institution of a new health code regulation banning artificial trans fat from New York City restaurants (Angell et al. 2009). This study showed that educational persuasion of restaurants failed, but with public support, the regulatory approach succeeded in reducing the proportion of restaurants using trans fats from 50% in 2005 and 2006 to 1.6% in 2008. This project may be seen as a success predicated upon antecedent generation of public support, coupled with a legally enforceable regulation, analogous to the tobacco control successes cited in Sect. 11.4.6 above. The NYC study is an example of a single community (NYC) with a set of five serial surveys before, during, and after the intervention, without a comparison control community. Although causal inference is threatened by such a design, this risk was outweighed by both the repeated measurements and the finding of an extraordinarily large effect from the regulatory intervention (Angell et al. 2009).

TCS, the first, and PAT and NYC trans fat, the fourth and fifth examples, are opposites: TCS had a maximum of mass media education and a minimum of organizing, whereas PAT had the opposite. NYC had the benefit of antecedent nationwide education over the dangers of heart disease from trans fat, with excellent success from organizing an enforceable regulation. Thus, any of these three

[1] The Stanford Prevention Research Center collaborated with the investigators of this study.

models work, but to change complex behaviors, community educators must use sophisticated behavior-change methods, such as those of the health communication-behavior change theory described above. If fear of arrest or a fine for breaking existing laws and regulations suffice to change personal drinking-and-driving habits or restaurant practices, then the education process is simpler.

11.6 Past Experience Leads to a "Master Plan"

A five-step approach emerges from these studies, with many lessons derived from the Stanford FCP done in the early 1980s:

(a) Define the problem. Does the community need a health promotion campaign? A decision can be made from national or local survey data.

(b) Organize the community, creating a campaign that includes behavior change education, continued organizing, training of community organizations, empowerment of the public and the community's organizations, and future maintenance of programs developed during an initial phase of about 2 years.

(c) Implement an intervention, delivering 3–5 h of education exposure directed toward the total adult population per year for 2 years, using in year one, about 70% television/radio to arouse interest and provide knowledge, about 15% from more specific and instructive print media (emphasizing newspapers, if available) and 15% from community events and programs (such as health fairs, contests, and classes). In the second year, decrease TV/radio to about 40% and increase print to 35% and community events and programs to 25%.

(d) Evaluate the intervention, using the process and summative methods described above. Insure that adequate tracking of exposure to the intervention (the independent variable) is done, with calculations of both the amount and type of intervention. Preferably use a repeated survey cohort design in order to obtain a clear measure of the intervention's effect on behavior and risk. If resources allow, add a repeated independent sample design for a better measure of the total community impact on the community, recognizing the dilution of effect from in-migration. As an example, 33% of the FCP's final independent samples were relatively new residents. The independent sample design, with its greater variance and inclusion of migrants, can therefore miss the important question of whether the intervention is working.

(e) Institutionalize programs. The community becomes a demonstration project, with its "empowered" organizations functioning as health promotion resource centers for a wider region.

(f) Use the new community resources and its residents' potential power to advocate for local, regional, and national governmental regulations and laws that will increase local intervention effects and extend them beyond the community.

Any individual or organization that wishes to engage in community-based health education should gain courage from the words of the now deceased anthropologist, Margaret Mead, as quoted in the closing passage of the The Catalonia Declaration

(1996, p. 75), “Never doubt the capacity of a few dedicated individuals to change the world, in fact, it is the only way it ever has.”

11.7 The Cultural Basis for Barriers to Success

A *healthy city* has been defined by the absence of crime, crowding, and poverty and by the presence of educated residents and enlightened (and trained) organizations. Together these lead to a community empowered to solve its social problems (i.e., to increase its social capital) (Travers 1997; Smith 2007).

Considering barriers, MacIntyre (2000) found Glasgow’s environmental factors to be major barriers to healthful exercise behavior. Certain macroeconomic factors inherent in *globalization*, such as capital flight and increased wealth and income gaps, have been described as barriers to planned change (Cahill 1983; Bezruchka 2000). All such barriers threaten *community stability*, inhibiting greatly the success of health promotion attempts. However, wise compensatory resource allocation can overcome many barriers, as was shown in the Hispanic minority of the TCS. Therefore, the challenge for the future is: Responsibility for success in community-based health interventions lies with the interventionist, not with the community’s residents! Secondly, it is clear that successful community-based health promotion requires attention to the cultural, environmental and social determinants of health that underlie the modern epidemics of chronic disease. Colditz (2001) reminds us of Rudolph LK Virchow’s words, to paraphrase, that: “The history of epidemics is therefore the history of disturbances of human culture.” Certainly the increasing worldwide prevalence of obesity, the spread of tobacco use, and the urban barriers to physical activity fit Virchow’s definition of “disturbances of human culture.” As Colditz points out, many scientific endeavors put false hope in the search for new risk factors, rather than in applying the now well-established means and economic benefits of reducing the culturally determined risk factors for chronic disease. For example, Aldana (2001) cites evidence for cost savings from three to six times the cost of health promotion for lifestyle change in large worksites. Despite this and other extensive evidence for the benefits of primary prevention (The Victoria Declaration on Heart Health 1992; The Catalonia Declaration 1996; IOM 2010), many countries, including the USA, spend less than 5% of total health care expenditures on prevention activities of all kinds (including immunizations, mammography, and health promotion) (Marwick 1994; World Health Report 1997). Although the reasons for the *unreasonable* diversion of a nation’s resources from disease prevention to disease treatment are many, it certainly includes pressure exerted by the pharmaceutical and medical device industries. Also, the tobacco industry is still a prosperous growth industry worldwide, especially in middle- and low-income countries (The Osaka Declaration 2001; Sebrie and Glantz 2006; IOM 2010). Therefore, from a political economy perspective, success in chronic disease prevention requires vigorous governmental policy changes (i.e., on taxation, advertising, promotion of tobacco use to minors, etc.), in addition to more economic resources (The Osaka Declaration 2001; Koh et al. 2007; IOM 2010).

11.8 Conclusions

Four decades of the "total community" health promotion approach in developed countries strongly support the feasibility, at relatively low cost, of achieving transfer of public education technologies to a community's infrastructure (public health, media, schools, etc.), resulting in significant changes in health habits of populations. Although most studies derive from small communities, recent successes in Tianjin (China), a city of 400,000 (two exemplars examined populations of >100,000), suggest that the model also works in large populations (Schooler et al. 1997). Theory matters: When the population gains self-efficacy through education, the result – *community efficacy* – enhances capacity to change institutional policy and practice, thus maintaining community change. Organizing and educating communities requires *advocacy*, *activism*, *coalition building*, and *leadership*; success is enhanced by *regulatory change*. Synergism occurs when the needed laws and regulations come from state, provincial or national governments. Such dual approaches are needed for any major success in preventing chronic disease, which is particularly important for developing countries (IOM 2010).

Science cannot serve society if its evidence for educational benefit is ignored. This is not a new concept. As written over 2,000 years ago in the Chinese *Book of Lessons,* "If a virtuous and learned scholar aims to influence the people as a whole, one must first educate the people."

References

Aldana S (2001) Financial impact of health promotion programs: a comprehensive review of the literature. Am J Health Promot 15:296–320

Altman DG, Flora JA, Fortmann SP, Farquhar JW (1987) The cost-effectiveness of three smoking cessation programs. Am J Public Health 77:162–165

Angell SY, Silver LD, Goldstein GP, Johnson CM, Deitcher DR, Frieden TR, Bassett MT (2009) Cholesterol control beyond the clinic: New York City's trans fat restriction. Ann Intern Med 151:129–134

Baker L, Wagner TH, Singer S, Bundorf MK (2003) Use of the internet and e-mail for health care information: results from a national survey. JAMA 289:2400–2406

Bandura A (1986) Social foundations of thought and action: a social cognitive theory. Prentice-Hall, Englewood Cliffs

Bezruchka S (2000) Is globalization dangerous to our health? West J Med 172:332–334

Cahill J (1983) Structural characteristics of the macroeconomy and mental health: implications for primary prevention research. Am J Community Psychol 11:553–571

Cartwright D (1949) Some principles of mass persuasion. Hum Relat 2:253–267

Cirksena MK, Flora JA (1995) Audience segmentation in worksite health promotion: a procedure using social marketing concepts. Health Educ Res 10(2):211–224

Colditz GA (2001) Cancer culture: epidemics, human behavior, and the dubious search for new risk factors. Am J Public Health 91:357–359

Farquhar JW, Maccoby N, Wood P, Alexander JK, Breitrose H, Brown BW, Haskell WL, McAlister AL, Meyer AJ, Nash JD, Stern MP (1977) Community education for cardiovascular health. Lancet 1:1192–1195

Farquhar JW, Fortmann SP, Flora JA, Taylor CB, Haskell WL, Williams PT, Maccoby N, Wood PD (1990) The Stanford Five-City Project: effects of community-wide education on cardiovascular disease risk factors. JAMA 264:359–365

Farquhar JW, Fortmann SP, Flora JA, Maccoby N (1991) Methods of communication to influence behaviour. In: Holland WW, Detels R, Knox G (eds) Oxford textbook of public health, vol 2, 2nd edn. Oxford University Press, Oxford

Fiore MC, Jaen CR (2008) A clinical blueprint to accelerate the elimination of tobacco use. JAMA 299:2083–2085

Flora JA (1991) Integrating evaluation into the development and design of risk communication programs. In: Fisher A, Pavlova M, Covello V (eds) Evaluation and effective risk communications workshop proceedings. U.S. Government Printing Office, Washington, DC, pp 33–40

Flora JA, Maccoby N, Farquhar JW (1989) Communication campaigns to prevent cardiovascular disease: the Stanford studies. In: Rice R, Atkin C (eds) Public communication campaigns. Sage, Beverly Hills, pp 233–252

Flora JA, Saphir MN, Schooler C, Rimal RN (1997) Toward a framework for intervention channels: reach, involvement and impact. Ann Epidemiol S7:S104–S112

Fortmann SP, Flora JA, Winkleby MA, Schooler C, Taylor CB, Farquhar JW (1995) Community intervention trials: reflections on the Stanford Five-City Project. Am J Epidemiol 142(6): 576–586

Green LW, Kreuter MW (1991) Health promotion planning: an educational and environmental approach, 2nd edn. Mayfield Publishing, Mountain View

Hedeker D, Gibbons RD, Flay BR (1994) Random-effects regression models for clustered data with an example from smoking prevention research. J Consult Clin Psychol 62:757–765

Holder HD, Wallack L (1986) Contemporary perspectives for prevention of alcohol problems: an empirically-derived model. J Public Health Policy 7:324–330

Holder HD, Salz RF, Grube JW, Treno AJ, Reynolds RI, Voas RB, Gruenewald PJ (1997) Summing up: lessons from a comprehensive community prevention trial. Addiction 92(S2):S293–S301

IOM (Institute of Medicine) (2010) Promoting cardiovascular health in the developing world: a critical challenge to achieve global health. The National Academies Press, Washington, DC

King AC, Ahn DK, Oliveira BM, Atienza AA, Castro CM, Gardner CD (2008) Promoting physical activity through hand-held computer technology. Am J Prev Med 34:138–142

Koh HK, Joosens SX, Connolly GN (2007) Making smoking history worldwide. NEJM 356: 1496–1498

Kotler P (1975) Marketing for nonprofit organizations. Prentice Hall, Englewood Cliffs

Lefebvre RC, Flora JA (1988) Social marketing and public health interventions. Health Educ Q 15:299–315

MacIntyre S (2000) The social patterning of exercise behaviours: the role of personal and local resources. Br J Sports Med 34:6

Marwick C (1994) New disease prevention effort goes by the book. JAMA 272:993–994

NHS Information Service (2007) Statistics on smoking, England

Pierce JP, Evans N, Farkas NJ (1994) Tobacco use in California. An evaluation of the Tobacco Control Program 1989–1993. University of California, La Jolla/San Diego

Puska P (1995) Experience with major subprogrammes and examples of innovative interventions. In: Puska P, Nissinen A, Vartianinen E (eds) The North Karelia Project. Helsinki University Printing House, Helsinki, pp 159–167

Puska P, Nissinen A, Tuomilehto J, Salonen JT, Koskela K, McAlister A, Kottke TE, Maccoby N, Farquhar JW (1985) The community-based strategy to prevent coronary heart disease: conclusions from the ten years of the North Karelia project. Annu Rev Public Health 6:147–193

Roberts DF, Maccoby N (1973) Information processing and persuasion: counterarguing behavior. In: Kline F, Clarke P (eds) Sage communication research annuals, vol II. Sage, Beverly Hills

Robinson TN, Killen JD, Taylor CB, Telch MJ, Bryson SW, Saylor KE, Maron DJ, Maccoby N, Farquhar JW (1987) Perspectives on adolescent substance abuse: a defined population study. JAMA 258:2027–2076

Rogers E (1983) Diffusion of innovations. Free Press, New York

Schooler C, Farquhar JW, Fortmann SP, Flora JA (1997) Synthesis of findings and issues from community prevention trials. Ann Epidemiol S7:S54–S68

Schooler C, Chaffee SH, Flora JA, Roser C (1998) Health campaign channels: tradeoffs among reach, specificity, and impact. Hum Commun Res 24:410–432

Sebrie E, Glantz S (2006) The tobacco industry in developing countries. BMJ 332:313–334

Slater MD, Flora JA (1991) Health lifestyles: audience segmentation analysis for public health interventions. Health Educ Q 18(2):221–233

Smith MK (2007) Social capital. The encyclopedia of informal education. http://www.infed.org/biblio/social_capital.htm

The Catalonia Declaration: Investing in Heart Health (1996) Declaration of the Advisory Board, Second International Heart Health Conference 1995. Department of Health and Social Security of the Autonomous Government of Catalonia, Barcelona

The Chinese Classics: The Great Learning, one of four books written 2000 years ago

The Osaka Declaration (2001) Health, economics and political action: stemming the global tide of cardiovascular disease. Declaration of the policy board. The Fourth International Heart Health Conference, Osaka

Travers KD (1997) Reducing inequities through participatory research and community empowerment. Health Educ Behav 24:344–356

Truett J, Cornfield J, Kannel E (1967) Multivariate analysis of the risk of coronary heart disease in Framingham. J Chronic Dis 20:511–524

Victoria Declaration on Heart Health (1992) Declaration of the advisory board, first international heart health conference. Health and Welfare Canada, Victoria

Victorian Health Promotion Foundation (1994) Annual report, 1994. VicHealth, Victoria

Williams J, Flora JA (1995) Health behavior segmentation and campaign planning to reduce cardiovascular disease risk among Hispanics. Health Educ Q 22(1):36–48

Williams PT, Fortmann SP, Farquhar JW, Varady A, Mellen S (1981) A comparison of statistical methods for evaluating risk factor changes in community-based studies: an example from the Stanford Three Community Study. J Chronic Disease 34:565–571

World Health Organization (1986) Report of a WHO Expert Committee. Community prevention and control of cardiovascular disease. Technical report series No. 732. World Health Organization, Geneva

World Health Report (1997) Conquering suffering, enriching humanity. World Health Organization, Geneva

Internet-Based Epidemiology 12

Lorenzo Richiardi, Costanza Pizzi, and Daniela Paolotti

Contents

L. Richiardi (✉)
Unit of Cancer Epidemiology and Centre for Oncologic Prevention, Department of Medical Sciences, University of Turin, Turin, Italy

C. Pizzi
Cancer Epidemiology Unit, Department of Medical Sciences, CPO-Piemonte and University of Turin, Turin, Italy

Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK

D. Paolotti
Computational Epidemiology Lab, ISI Foundation, Turin, Italy

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_59,

12.1 Introduction

Currently, almost two billion persons worldwide, that is 30% of the world population, have access to the Internet (Internet World Stats). These numbers are accurate as of June 30, 2010, and will soon be outdated: access has grown by 445% from 2000. The proportion of users varies by country, being 77% in North America and 58% in Europe, with notable peaks in the Nordic countries, where Denmark and Finland are above 85% and Norway and Sweden are above 90%. Figure 12.1 provides a global view on the proportion of users and growth in use in the last 10 years.

The proportion of users is not homogenously distributed in the population. In 2007, for example, according to Eurostat data, in all EU countries, with no exceptions, the proportion of Internet users was higher among men than women and among people aged 16–24 than older persons (United Nations Economic Commission for Europe (UNECE) 2011). It is reasonable, however, to expect that sooner or later almost everybody in the world will have access to the Internet, with no marked country, age, or sex differences. This is a very attractive prospect for epidemiologists and human researchers in general, who already recognized the possibility of using the Internet to conduct field studies in the early 1990s.

It is perhaps surprising that the Internet has been initially used mainly to conduct surveys rather than longitudinal studies or interventions, although the latter are less vulnerable to selection bias. The first medical surveys, such as those of patients with inflammatory bowel diseases (Hilsden et al. 1999; Soetikno et al. 1997) or diabetes (Baehring et al. 1997), have been published at the end of the 1990s. These

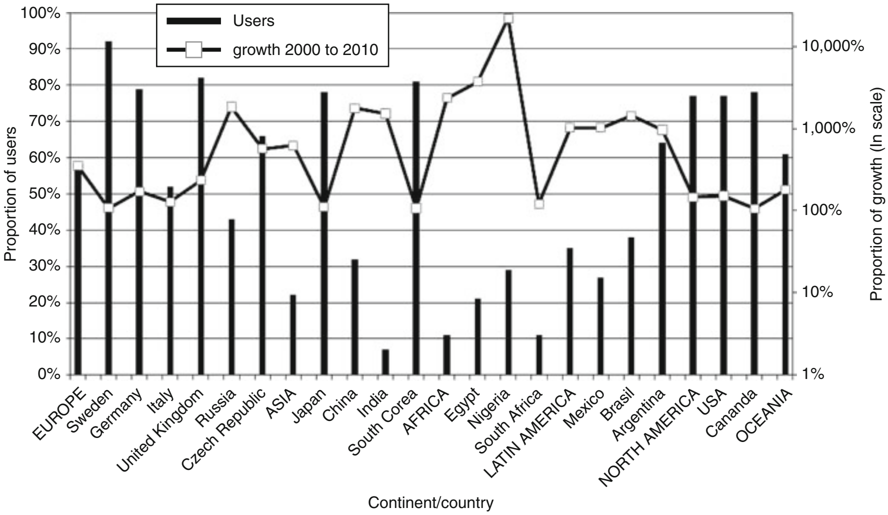


Fig. 12.1 Proportion of Internet users on June 30, 2010, and growth from 2000 to 2010 in selected countries (Internet World Stats 2011)

studies were pioneered by surveys carried out by psychologists and sociologists (Berrens et al. 2003; Buchanan and Smith 1999; Kraut et al. 2004; Skitka and Sargis 2006). In 1997, Kushi and colleagues reported the launch of a pilot study for an Internet-based cohort on diet and breast cancer (Kushi et al. 1997), and a small number of web-based birth cohort studies have been conducted in the last 10 years (Hercberg et al. 2010; Mikkelsen et al. 2009; Richiardi et al. 2007; Treadwell et al. 1999; Turner et al. 2009). However, most of the Internet-based medical studies are currently intervention trials. The Internet was first suggested as a tool to manage all aspects of the trial, including randomization and data acquisition (Kelly and Oldham 1997; Pepine et al. 1998), but, in the last 10 years, it has often been used for the purpose of recruiting study participants (McAlindon et al. 2003; Wang and Etter 2004).

The idea of using the Internet in empirical research in general and epidemiological research in particular often receives skeptical reactions. Typical concerns include problems related to lack of exposure heterogeneity in the study participants, phantom participations, duplicate records by the same participant, low data quality, and confidentiality issues (Gosling et al. 2004; Skitka and Sargis 2006; van Gelder et al. 2010). In fact, most of these problems have limited implications or can be solved technically. For example, the lack of heterogeneity does not apply to most of the exposures, as in many countries more than half of the population has access to the Internet. Similarly, various methods have been developed to identify duplicate entries, to enhance data quality, and to ensure data security and confidentiality, including the use of encrypted connections, registration through individual username and password, the use of screening questions to detect duplicates, and checks for implausible answers (Baer et al. 2002; Bowen et al. 2008; Dillman and Smyth 2007; Gosling et al. 2004; van Gelder et al. 2010).

As for any study based on volunteers, the main critical issue related to Internet-based research is, however, the representativeness of participants for the study population and the likelihood of selection bias. This is currently largely debated. At an international conference held in 2008, for example, one of the authors heard the editor of an epidemiological journal saying that because of concerns about self-selection, he was a priori against publishing results of Internet-based surveys. Issues of selection and selection bias will be extensively discussed in this chapter both in an ad-hoc section and within the context of the discussion about the use of the Internet in each type of study design.

In contrast to its limitations, the use of the Internet in epidemiological research offers several advantages, including decreased costs, simplified logistics, rapidity, flexibility, the possibility to tailor the questionnaire to the participants' characteristics, and instantaneous checks to identify inconsistencies as well as to reduce errors resulting from data entry (Baer et al. 2002; Dillman and Smyth 2007; van Gelder et al. 2010). Moreover, online studies have fewer constraints than traditional studies, both from a geographical and a temporal point of view: they can reach distant or "hidden" populations as well as they can recruit continuously for several years. Some other advantages are specific to the different study designs and will be discussed further in the corresponding sections of this chapter. However, a common feature of Internet-based studies, often overlooked, is the active involvement and empowerment of the study subjects (Rhodes et al. 2003). They can give feedback over the whole duration of the project and receive (and comment on) information about the study results. For example, it is not uncommon that, within the framework of an Internet-based study, researchers keep constant contact with the study participants using social networks.

In several studies, the Internet has been used to submit online questionnaires to a prespecified population, including members of a "traditional" cohort (Ekman et al. 2006; Russell et al. 2010), members of Internet panels (Silver et al. 2002; West et al. 2006), or members of a mailing list (Ruddy et al. 2010). In all these studies, the Internet was not used to directly recruit participants, whereas it was used as a tool to deliver a questionnaire to a preselected population. We will not discuss advantages and limitations of online questionnaires and technicalities on how to prepare them in this chapter (about these issues, see, for example, references Baer et al. 2002; Dillman et al. 2009; Ekman et al. 2007; Kongsved et al. 2007; Russell et al. 2010; Schleyer and Forrest 2000; van Gelder et al. 2010). Rather, we will focus on the Internet as a method to recruit study participants and its influence on study design and validity. Typically, Internet-based studies use online questionnaires, but this is not a necessary feature. Indeed, researchers may have a direct contact with participants via the Internet, for example, to complete a telephone interview or to obtain biological samples (Etter et al. 2005; Richiardi et al. 2007).

In this chapter, we will address general methodological issues about Internet-based studies and discuss examples of the use of the Internet in the context of different types of epidemiological designs, from surveys to randomized studies. Debates about the use of the Internet in epidemiological research may be affected by preconceived opinions either in favor or against it. Being involved in Internet-based research, we cannot be objective, but we will aim at discussing the different issues impartially.

12.2 Selection and Selection Bias in Internet-Based Studies

Participants in an Internet-based study are doubtless a selected population, regardless of the epidemiological design chosen to carry out the study. This is due to two main reasons: (1) the source population is restricted to Internet users (either all Internet users or users of specific websites), and (2) participation of subjects should be voluntary. What does instead depend on the study design is the mechanism through which bias may be induced by the sample selection process. We will illustrate these mechanisms in different study designs also using directed acyclic graphs (DAG) (see chapter ▶Directed Acyclic Graphs of this handbook) indicating E as the exposure of interest, R as a risk factor (or a set of risk factors) for the outcome of interest D, U_i as other unknown/unmeasured variables, and S as the indicator of selection into the sample. In all figures, a square around S will indicate conditioning on study participation.

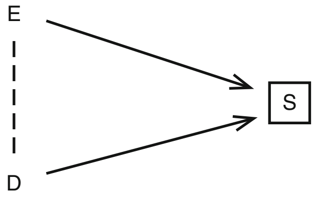


Fig. 12.2 Diagram of a study where selection of the study subjects (*S*) depends on both the exposure (*E*) and the outcome (*D*) of interest. The *dashed line* represents an association induced by conditioning on *S* (represented by a square around *S*)

In surveys aiming at estimating the prevalence of a disease, there is selection bias if the disease status or any determinant of the disease is associated with the selection probability. This is very likely to happen in Internet-based studies. If determinants of selection are known, it is possible to apply weights and obtain a valid estimate of the prevalence. However, in most situations, it is very difficult to obtain a good picture of the determinants involved in the selection in a study based on volunteers recruited through the Internet.

In studies aiming at estimating associations, in general there is selection bias if the probability of selection depends on both the exposure (E) and the outcome (D) of interest (Hernán et al. 2004). This is illustrated in Fig. 12.2. For example, in an Internet-based cross-sectional study on asthma, the probability of volunteering to take part into the study could be associated with having asthma (D) as well as with living in a heavily polluted area (E). Under these circumstances, because of the conditioning on selection, an Internet-based study would most likely find an association between air pollution and asthma even in the case of a lack of a true causal association between these two variables. This type of selection bias can be large.

12.2.1 Longitudinal Studies

In longitudinal studies where the outcome of interest occurs after being selected into the study sample, the mechanism that we have just described does not apply but selection bias may still occur (Glymour 2006; Hernán et al. 2004). In particular, if the likelihood of participation in the study depends on both exposure (E) and a disease risk factor (R) for the outcome of interest (D), and these are independent of each other in the general population, the selection process induces an association between E and R; R becomes a confounder of the E-D association and thus, if we cannot adjust for it in the analyses, the estimate of the association is biased (Fig. 12.3). We (Pizzi et al. 2011) and others (Greenland 2003) showed that the potential bias induced by this mechanism is usually moderate. For example, assuming no effect of the exposure on the incidence of the disease (true rate ratio = 1.0), the estimated exposure to disease rate ratio is 0.95 even if (1) the odds ratio of selection associated with the exposure is 2.0, (2) the odds ratio of selection

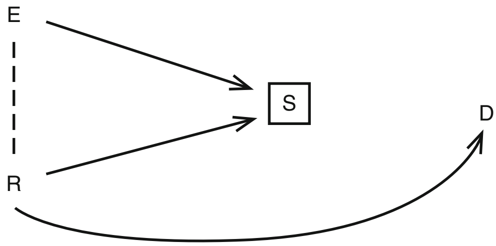


Fig. 12.3 Diagram of a cohort study based on a selected sample. In the population, the exposure of interest (*E*) is independent of the risk factor (*R*) and is not associated with the outcome of interest (*D*). Both *E* and *R* affect the likelihood of being selected as member of the study (*S*). In the selected sample (i.e., conditioning on *S* – represented by a square around *S*), *E* and *R* become associated (indicated by a *dashed line*)

associated with the risk factor is 4.0, and (3) the risk factor increases the risk of the disease by fourfold (Pizzi et al. 2011). The ideal situation is a randomized study in which, thanks to randomization which occurs after selection of study subjects, the exposure is not associated with selection in the study. Under this situation, assuming compliance to the treatment assigned, there is no possibility for selection bias even when some risk factors (R) are strong determinants of the selection.

It should be noted that, when the exposure and the disease risk factor are associated in the general population, the selection mechanism will alter the confounding effect of the risk factor, either increasing or decreasing it, according to the strength and direction of the association between E and R (in the general population) and also of the associations between E and R with S.

Let us now focus on the specific case of an Internet-based *cohort study* and the effect of restriction on Internet users, by discussing some hypothetical examples.

1. In the first example, the exposure of interest (E) and the risk factor (R) are independent in the general population and both are associated, either directly (Fig. 12.4a) or indirectly (Fig. 12.4b), with the likelihood of being an Internet user and thus of being selected into the sample.
   In Fig. 12.4a, for example, socioeconomic status (E) as well as year of birth (R), which are assumed to be independent in the general population, could affect the probability of being an Internet user. In Fig. 12.4b, socioeconomic status (U_1) could be a cause of being an Internet user and of being a smoker (E), while year of birth (U_2) could affect both the likelihood of being an Internet user and height (R). In these scenarios, the restriction on Internet users induces a spurious association between E and R, and, therefore, year of birth becomes associated with socioeconomic status (Fig. 12.4a) and smoking becomes associated with height (Fig. 12.4b).

2. In a second example, the exposure of interest (E) and the disease risk factor (R) are already associated in the general population as they share a common cause (U_3). This scenario is illustrated in Fig. 12.5.

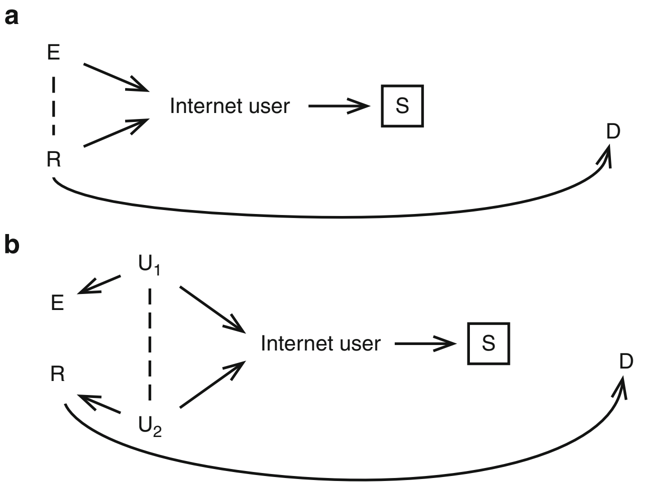


Fig. 12.4 Diagram of an Internet-based cohort study. (**a**) In the population, the exposure (*E*) and the disease risk factor (*R*) are independent, and *E* is not associated with the outcome of interest (*D*). Both *E* and *R* affect the likelihood of being an Internet user and thus of being selected as member of the study (*S*). In the selected sample (i.e., conditioning on *S* – represented by a square around *S*), *E* and *R* become associated (indicated by a *dashed line*). (**b**) Here, *E* and *R* affect the likelihood of being an Internet user as proxy of other factors (U_1 and U_2)

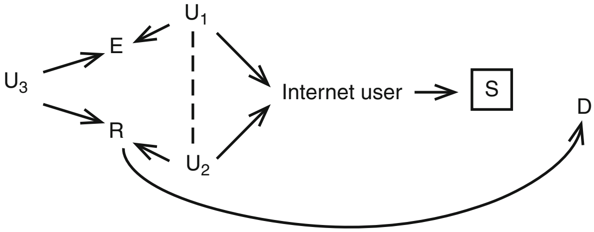


Fig. 12.5 Diagram of an Internet-based cohort study. In the population, the exposure (*E*) and the disease risk factor (*R*) are associated as they share a common cause U_3, and both *E* and *R* are associated with the likelihood of being an Internet user. In the selected sample (i.e., conditioning on *S* – represented by a square around *S*), U_1 and U_2 become associated (indicated by a *dashed line*) thus altering the original association between *E* and *R*

This is in fact an extension of the scenario depicted in Fig. 12.4b, where an additional factor (U_3), for example, place of birth, could affect both smoking (E) and height (R). Under this scenario, it is hard to predict whether the restriction of the source population to Internet users would increase or decrease the bias, as this depends on the strength and direction of the E-R association existing in the general population and on the strength and direction of the spurious E-R association induced by the sample selection process (as a consequence of the association induced between U_1 and U_2).

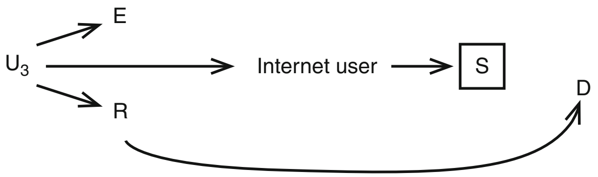


Fig. 12.6 Diagram of an Internet-based cohort study. In the population, the exposure (*E*) and the disease risk factor (*R*) are associated as they share a common cause U_3, which also affects the likelihood of being an Internet user. In the sample, the condition on *S* (represented by a square around *S*) implies a partial condition on $U_{3,}$ thus attenuating the association between *E* and *R*

3. A further example of interest is when E and R are associated in the general population because they share a common cause which, in turn, is a determinant of being an Internet user. This is shown in Fig. 12.6 that summarizes the case of a study in which socioeconomic status (U_3) is a cause of being an Internet user as well as of being a smoker (E) and taking regular exercise (R).

Under this scenario, restricting the study to Internet users implies a partial conditioning on U_3 and therefore a likely attenuation of the E-R association among the study participants compared with the general population. This means that the restriction would diminish the confounding effect of R and thus the corresponding bias induced in the exposure-outcome association.

In these examples, we only considered the restriction to Internet users as a potential source of selection. Similar considerations can be made for the second source of selection that is volunteering. Our examples demonstrate, however, that the effect of selection on the exposure-disease association in longitudinal studies is difficult to predict and can induce or attenuate confounding bias present in the general population (Pizzi et al. 2012). However, even when it induces bias, its magnitude is expected to be small as discussed above.

If the aim of the study is to estimate the incidence of the disease, bias is more likely to occur. The disease status itself cannot affect the likelihood of selection (as it may occur in surveys) but any disease risk factor (R in Figs. 12.3–12.6) that is associated with being an Internet user or volunteering would bias the estimate of the incidence.

12.2.2 Case-Control Studies

Case-control studies can be restricted to a specific population and still reveal valid associations (apart from some of the considerations just mentioned for cohort studies). Thus, the restriction to cases who are Internet users (or, e.g., who seek information about their disease on the Internet) is not expected to introduce a large bias by itself. The concern is, however, on the effect of this restriction on the control selection, as control subjects should be representative of the source population from which the cases originated. When cases and controls arise from two different source

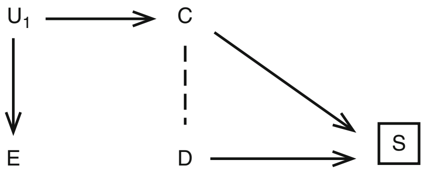


Fig. 12.7 Diagram of a case-control study in which cases have the disease of interest (*D*) and controls have a control disease (*C*). Both *C* and *D* affect the probability of selection in the study (*S*), and *C* is also associated with the exposure of interest (*E*), as they share a common cause U_1. In the selected sample (i.e., conditioning on S - represented by a square around S), C and D become associated (indicated by a dashed line) thus altering the original association between E and D

populations, there are clearly factors with different distributions across the two groups, which, when associated with the exposure of interest, lead to cases and controls being no longer comparable.

Even when the principle of the same source population is met, selection bias may occur if the exposure is associated, either directly or through other factors, with the probability of selection, that is, the sample fraction of controls (and cases when applicable) is not constant across exposure levels. In this situation, the causal structure becomes similar to that depicted in Fig. 12.2 or Fig. 12.7, both implying selection bias. Figure 12.7, for example, could depict a scenario in which cases with a disease of interest (D) are compared with controls originating from the same source population but having another disease (C). If the disease (C) is associated with the exposure of interest (E) (in this scenario through the common cause U_1), then there is selection bias (Hernán et al. 2004).

Since participants in an Internet-based case-control study are self-selected, it is not unlikely that some of the determinants of the exposure or the exposure itself are associated with volunteering and thus with selection. It should be noted, however, that the causal structures summarized in Figs. 12.2 and 12.7 do not always introduce bias. Indeed, there is no bias if the ratio of the selection probabilities of exposed and unexposed cases is the same as the ratio of the corresponding selection probabilities in exposed and unexposed controls. Under this special scenario, E and D are still determinants of the selection but the association remains to be valid.

12.3 Internet-Based Recruitment

The use of the Internet allows recruitment of study participants from large populations with a decrease in cost and time. Internet recruitment is similar to “traditional” studies in that it requires the specification of the source population. For example, the source population of an Internet-based study aiming at surveying cancer patients residing in Italy is defined by those Internet users who reside in Italy and have cancer. However, in Internet-based studies, there is no available list for random sampling from the source population and participants are self-selected volunteers. Thus, apart from defining the source population by means of eligibility criteria,

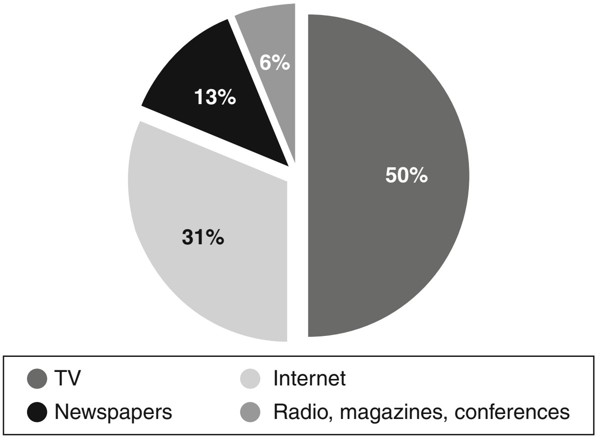


Fig. 12.8 Distribution on how participants to the Influweb study got to know about the existence of the study (based on about 2,000 respondents), Influweb, Italy, 2008–2009 (Influenzanet 2011)

researchers have little possibility to influence the selection of study participants. Internet-based recruitment typically goes through two processes: (1) people in the source population need to become aware of the existence of the study, and (2) they need to agree to participate.

The study has to be advertised in order to efficiently reach all members of the source population. The advertisement of a study of individuals with a specific disease or exposure would have to involve websites targeting that disease/exposure, while a study targeting the general population would be more efficiently advertised by means of a publicity campaign involving television, radio, and newspaper coverage and word of mouth. The Internet-based Nutrinet-Santé study, for example, which aims at recruiting a large cohort of individuals from the French general population, advertises the existence of the study through a multimedia campaign, different websites, and professional health channels (Hercberg et al. 2010). On the basis of our experience in the Italian Influweb study (an Internet-based monitoring project for influenza surveillance in Italy), when the target is the general population, television seems to be the most effective means of communication, especially if the advertisement takes place in Scientific Communication programs, while articles on the front page of online newspapers seem to be very efficient in terms of visits to the website but not in terms of new volunteers (Fig. 12.8) (Paolotti et al. 2010).

The Danish Pregnancy Planning Study, an Internet-based cohort study of women of reproductive age, was advertised through a pop-up advertisement in a national health-related website as well as with press releases to reach the media (Mikkelsen et al. 2009). An Internet-based study of determinants of gout attacks was advertised using a Google advertisement linked to the search term “gout” (Zhang et al. 2007). Over a period of 11 months, the advertisement was displayed 866,703 times, 6.6% of which led to a visit of the study website.

An often overlooked issue is that an integrated approach to advertise an Internet-based study aiming at recruiting for a long period implies intense and continuous efforts. Conversely, if the study is only advertised via selected websites, the recruitment process is less demanding, although the recruitment rate will be lower. Thus, to reach large sample sizes, an integrated approach is generally necessary. Furthermore, the methods used to advertise the existence of the study will impact on the characteristics of the study participants and the probability of selection bias. If a mother-child cohort, for example, is advertised only during antenatal courses, there will be an oversampling of nulliparous women who are more likely to attend these courses. An integrated approach could reduce this source of selection and spread the information about the study to the whole source population. A recent study compared three methods of advertisement for a survey of young adult smokers, namely, (1) advertisement on a single general website (Craiglist 2011), (2) an Internet campaign to target social networks and lifestyle-based websites, and (3) an invitation sent to members of an Internet panel (Ramo et al. 2010). The Internet campaign yielded the largest number of participants, but it was less cost-effective (about US$43 per completed survey) and was the method associated with the largest proportion of incomplete surveys. Roughly spoken, the three samples differed in terms of age, sex, ethnicity, education attained, nicotine dependence, and recent use of marijuana and cigars. Fewer or no differences were found in other variables, including alcohol use and smoking prevalence.

Once the members of the source population are aware of the study, they should access the study website and volunteer to participate. This is obviously a key issue. During the pandemic season, from October 2009 to March 2010, the Influweb website (Influenzanet 2011), for example, was visited 90,000 times and roughly 3,000 persons participated. The success of volunteering depends on several elements: the study website should not only induce participation but the study topic should also be of interest among the source population. For example, influenza during the H1N1 pandemic was a topic prone to gain the interest of the public and to reach the media. In other words, Internet-based recruitment is more efficient when the study is about a topic of general interest (Paolotti et al. 2010; Tilston et al. 2010), or targets a strongly motivated population, such as smokers trying to quit smoking (Civljak et al. 2010) or pregnant women (Richiardi et al. 2007). Even when an integrated approach for advertisement is used and the population is motivated to participate in the study, the participation proportion is not going to be high compared to traditional studies. For example, we used several means to advertise the existence of the NINFEA (Nascita ed INFanzia: gli Eetti dell'Ambiente) study (NINFEA 2011), a birth cohort study, to the population of the city of Turin, Italy. We have recently estimated that about 3–4% of the total number of pregnant women present in the population (excluding those born outside Italy) participated in the study (Pizzi et al. 2012). Considering that about 60% of the pregnant women have access to the Internet, and thus belong to the source population, this proportion translates into a participation proportion of 5–7%.

Some Internet-based surveys collect anonymous data but many Internet-based studies require registration and collect demographic information. This is an obvious requirement for follow-up studies in which subjects should be identified and be recontacted, but it is also used in surveys, for example, to limit the problem of duplicate or phantom participations. Researchers undertaking a Web-based study should therefore check that their platform is compliant with privacy regulations in the country where the study is ongoing. Upon registration, users have to be informed what these requirements are and, as in any epidemiological study collecting non-anonymous data through questionnaires, they should sign an informed consent form. In many studies, an online informed consent is used to decrease costs and organizational efforts and to enhance participation (the alternative being a mailed hard copy of the informed consent to be signed by participants). For example, in the NINFEA study, women provide online consent when they register and complete the study questionnaires, while they provide an additional written consent if they also donate a saliva sample.

The use of the Internet to collect non-anonymous data may raise concerns about safety and confidentiality issues. Researchers typically use technical solutions, such as encryption, firewalls, HTTPS (hypertext transfer protocol secure) protocol etc., which provide the same level of safety as traditional epidemiological studies, as, for example, mail surveys (Baer et al. 2002; Kraut et al. 2004). It is possible that potential participants are reluctant to participate in an Internet-based questionnaire fearing that their data could be accessed by people from outside the study. To address this issue, within the NINFEA study, we interviewed a small number ($n = 37$) of women who were both aware of the study and had access to the Internet but did not participate in the study (Richiardi et al. 2007). No woman reported that she did not participate for fear of revealing personal information, while the most common reason of non-participation was lack of interest in the study.

12.4 Study Designs

The Internet has been used in the context of most of the classical epidemiological studies, including cross-sectional (see chapter ▸Descriptive Studies of this handbook), cohort studies (see chapter ▸Cohort Studies of this handbook), interventions (see chapter ▸Intervention Trials of this handbook), and case-control studies (see chapter ▸Case-Control Studies of this handbook). Basic characteristics of these study types are discussed in detail in the chapters listed above.

12.4.1 Cross-Sectional Studies

Cross-sectional studies or surveys are typically carried out to measure prevalence and are particularly vulnerable to selection bias. If participants have a different prevalence of the disease of interest compared to non-participants, and it is not possible to apply weights to counterbalance this difference, results of the study will be difficult to interpret. This drawback makes the use of the Internet in surveys very problematic.

Although characteristics of Internet users can be investigated and may sometimes be known for certain populations, determinants of self-selection in a specific study are almost unknown. It is possible to obtain a rough estimate of the number of individuals who visit the study website to estimate a sort of response proportion (number of participants out of the number of visitors). However, this proportion is relatively useless as the number of visitors differs from the number of subjects who became aware of the study. Furthermore, the response proportion would give little information on the amount of bias in the estimate of the disease prevalence, as the key issue is whether volunteering is associated or not with the probability of having the characteristic of interest.

Having this strong limitation in mind, Internet-based surveys can still be useful for a number of reasons, including to conduct qualitative studies in which population representativeness is less relevant, to rapidly obtain information to generate hypotheses or develop a study protocol, to reach hidden populations, and to identify patients with rare diseases.

Simmons and colleagues carried out an Internet-based cross-sectional study between 2001 and 2002 to obtain information on possible precipitating factors of multiple sclerosis (MS) (Simmons et al. 2004). The aim was to generate hypotheses and select potential precipitating factors to be investigated in a cohort study of MS patients. An anonymous questionnaire in English was posted on the MS Australia and the MS International Federation websites for a period of about 10 months. About 2,500 self-selected patients from 60 countries in total, mainly from the USA, Australia, and the UK, completed the questionnaire. They reported factors that in their opinion were improving their condition or worsening their MS symptoms.

Behavioral research or studies on HIV and sexually transmitted diseases among men who have sex with men (MSM) are most often conducted on convenience samples, as MSM is a hard-to-reach population (a so-called “hidden” population). Therefore, Internet-based surveys using anonymous questionnaires are becoming increasingly common in this field (Elford et al. 2009; Evans et al. 2007; Hirshfield et al. 2004; Rosser et al. 2009). A study conducted in London in 2002 and 2003 recruited about 4,000 participants from HIV-positive patients attending outpatient clinics (12%), men seeking an HIV test (10%), men using gyms in central London (35%), and Internet users, the latter either via chat rooms or the websites of gaydar (2011) and gay.com (2011) (43%). The samples had different socioeconomic and behavioral characteristics, and most likely none of them was representative of the whole MSM population.

A number of studies aimed at comparing characteristics of participants in an Internet-based cross-sectional study with study participants based on a representative sample of the population (Andersson et al. 2002; Etter and Perneger 2001; Evans et al. 2007; Klovning et al. 2009; Marcus et al. 2009; Miller et al. 2010; Ross et al. 2005). Unsurprisingly, participants recruited via the Internet had different characteristics and disease prevalence, most often in an unpredictable direction and magnitude. This reinforces the concept that cross-sectional studies can only be conducted over the Internet if their aim does not require a population representative sample.

12.4.2 Cohort Studies

In 1997, an editorial on the use of the Internet in epidemiology suggested the possibility of conducting a cohort study of Internet users, defining this study population as an epidemiologists' dream coming true (Rothman et al. 1997). Authors were briefly listing some of the most evident advantages, such as fast enrollment of a large sample, prolonged contact with cohort members, inexpensive and efficient follow-up, as well as some possible problems, underlying the risk of marauder and phantom users, and issues of information validity. According to a response letter to this editorial, Kushi and colleagues were already piloting at that time the feasibility of an Internet-based cohort study (Kushi et al. 1997).

Indeed, the use of the Internet to recruit a cohort is very attractive as, in longitudinal studies, a representative sample is not a necessary requirement to get valid associational estimates. The epidemiological dream, however, has infrequently materialized in the last decade (Hercberg et al. 2010; Mikkelsen et al. 2009; Richiardi et al. 2007; Treadwell et al. 1999; Turner et al. 2009).

In 2005, we started the NINFEA study, which is an Internet-based mother-child cohort carried out in Italy (NINFEA 2011). A parallel study was started 2 years later in New Zealand (The Early Life Factors Study of Childhood Diseases). We will therefore use the NINFEA cohort as an example to illustrate advantages and limitations of using the Internet to conduct a cohort study.

NINFEA is a multipurpose cohort aiming at investigating the effects of certain exposures during prenatal and early postnatal life on infant, child, and adult health (Richiardi et al. 2007). It enrolls pregnant women in order to follow-up their children for at least 18 years. Members of the cohort are children born to women who are Internet users, become aware of the study, and volunteer to participate. At any time during the pregnancy, they can register through the project website (www.progettoninfea.it) and complete the first questionnaire that lasts about 30 min. They are asked to complete two other 30-min long questionnaires at 6 and 18 months after delivery. Long-term follow-up involves linkage with available population registries and periodical very short online questionnaires focusing on specific outcomes (e.g., cognitive development, respiratory diseases).

We advertise the existence of the study using both "active" and "passive" methods. Active methods involve the collaboration of health personnel to distribute leaflets and/or to introduce the study to pregnant women when they reach hospitals or family clinics for reasons related to the pregnancy. Therefore, this approach is inherently limited to selected geographical areas and targets a (roughly) prespecified catchment population. Currently, the NINFEA study is actively advertised in the city of Turin (900,000 inhabitants), in the Tuscany Region (4,000,000 inhabitants) and, with a lower intensity, in the Piedmont Region (4,000,000 inhabitants including those living in Turin). One of the potential advantages of active recruitment that we have not yet explored is the involvement of specific populations characterized by high levels of specific exposure or diseases of interest. Let us suppose, for example, that in a community living around a large industrial area there are concerns on the possible reproductive effects of industrial emissions. It would be hard to quickly set up a "traditional" mother-child cohort in this population, especially if there are no research infrastructures already available in the area. It would, however, be possible to actively advertise the existence of the Internet-based cohort in that population in order to recruit a sufficient number of pregnant women. The online questionnaires could be modified accordingly to incorporate questions about the exposure of interest in the area.

Passive recruitment includes methods that do not involve the health personnel, including the Internet and the media. So far we have not launched a media campaign to advertise the NINFEA study, while we use the Internet in various ways: links to our study website posted on the hospitals' websites and on websites dedicated to pregnant women, participation in discussion forums related to pregnancy, and a NINFEA page in Facebook. This type of passive recruitment is not entirely automatic, as forums change constantly as well as they become more or less popular among Internet users. This implies constant monitoring of the accesses and the need to routinely post reminder messages. Furthermore, there should be bilateral interaction with the users of the forum to keep the discussion lively and attract new participants.

Doubtless, participants in the NINFEA cohort are strongly selected. When compared with the general population, we found that NINFEA participants have a higher socioeconomic status, a lower parity, are less frequently non-Italian citizens, and smoke less but have a higher alcohol consumption during pregnancy (Pizzi et al. 2012; Richiardi et al. 2007).

As mentioned above, participating women should complete 30-min follow-up questionnaires at 6 and 18 months after delivery and shorter questionnaires thereafter. The use of the Internet makes the follow-up rather efficient. In the NINFEA study, we collect information via e-mail, landline telephone, cell phone, and postal address at the time of the registration. When it is time to complete a follow-up questionnaire, we e-mail the women asking to access the website and complete the questionnaire. Non-responders are additionally contacted first by e-mail and then by telephone and regular mail. Currently, about 60–65% of the women reply after e-mail contacts, while remaining women have to be contacted using traditional approaches. Overall, the final response to the second and third questionnaires is about 85–90%.

During the first 5 years of the study, we have learned some valuable lessons regarding the follow-up. First, it is fundamental that contact information is obtained through mandatory questions at the time of the registration. This allows a much higher follow-up completeness at the cost of a small baseline dropout of participants who are not willing to reveal this type of information. Indeed, in cohort studies, baseline selection is a much smaller problem than incomplete follow-up, and the initial contact strategy should aim at assembling a cohort whose members guarantee high long-term participation. Second, although many authors have concerns about phantom participants in Internet-based studies, in the NINFEA cohort, this was a minor problem. Some people registered to the website to further understand about the questionnaires, but if public information about the project is clear enough, we believe that phantom participants and registration from non-eligible individuals are not important issues. For example, in our cohort, participation should occur before delivery. Indeed, so far, nobody participated after delivery, but we have been contacted by women asking if it is possible to participate after the baby was born. This suggests that the information on the website was clear enough to prevent registration after delivery. The possibility of duplicate registrations of the same participant is a more relevant issue. Although it is not difficult to identify them at the time of the statistical analyses, using key variables based on the available demographic information, duplicate registrations can make the follow-up procedures more complex. Let us take the example of a pregnant woman who registers twice with two different dates for the last menstruation (because of typos or because the pregnancy was redated between the two registrations): what date should be considered for the follow-up? When should the woman be recontacted? Most likely this person will be contacted twice. It is possible to introduce checks in order to limit the number of duplicate registrations and facilitate the follow-up procedures but, in our experience, it is not possible to avoid them completely. A third issue is change of e-mail address. Pregnant women and families with small children quite frequently change job and/or home. It is therefore important to keep frequent contact with the participants to give them updates about the study as well as to check contact information. If the e-mail address has changed, it is possible to contact the woman by telephone or letter and ask her to update her information. Indeed "traditional" cohort studies are affected by the same problem, and having the participants' e-mail address and a population restricted to Internet users only helps in obtaining a high follow-up participation proportion.

A potential limitation of the Internet-based recruitment is that there is no direct contact with the participants and thus, exposure information is self-reported and it is more difficult to obtain biological samples. In the NINFEA cohort, we collect samples of saliva from the mothers and the children using self-collection kits sent by regular mail. About 60% of the members of the cohort, so far, agreed to donate a sample. Successful collection of biological samples has also been achieved in other Internet-based studies, such as collection of saliva samples from subjects enrolled through a smoking cessation website (Etter et al. 2009). In an ongoing nationwide French cohort study aiming at recruiting half a million individuals (Nutri-Santeé cohort), participants can donate a blood sample by visiting local sample collection centers (Hercberg et al. 2010).

Currently the NINFEA study recruits and follows up about 25 subjects per week, employing overall (including IT experts and research assistants) less than three persons-years per year. Its Internet-based design offers two main advantages, namely, efficiency and flexibility, which are obtained at the cost of population representativeness. Flexibility is an often overlooked characteristic. The cohort will be able to recruit for an indefinite period, and its population coverage has changed and will change over time. Other Italian regions will be able to start active recruitment in the future, and it will be possible to adapt the questionnaire to other populations. Indeed, a parallel study has been launched in 2007 in New Zealand (The Early Life Factors Study of Childhood Diseases 2011), and other countries may join in the future. Furthermore, compared to traditional studies, it is easier to add, delete, or modify questions to all participants or selected subgroups, both to improve the questionnaire and to assess exposures previously uncovered. For example, during the H1N1 pandemic in the winter of 2010, we added some specific questions about vaccinations that were not planned in 2005. Based on our experience and other Internet-based cohort studies, we believe that this methodology will be increasingly used in the next years, especially when the target of the cohorts will concern highly motivated people (e.g., pregnant women), or populations difficult to reach or difficult to follow-up (e.g., short-term migrants) or when researchers will aim at fast recruitment of large samples.

12.4.3 Case-Control Studies

In case-control studies, cases and controls should be selected from the same source population, independently of their exposure status. As discussed before in this chapter, this is difficult to achieve using Internet-based recruitment. Thus, it is not surprising that we could not find any example of a study in which both cases and controls were selected using the Internet. A recent study describes the selection of a control sample to be used for genetic analyses: about 4,500 subjects were selected in the USA both among a group of Internet panelists and using ad-hoc Internet recruitment (Sanders et al. 2010). However, study cases were traditionally selected, and the study focused on genetic variants, which are unlikely to be strongly associated with self-selection in the study.

Case and control selection could be very problematic using the Internet but not impossible. We can imagine different approaches, which are described using a hypothetical example of a case-control study on celiac disease in Italy.

In a first approach, cases could be selected via a link posted on a specific website, for example, the website of the Italian Celiac Association. All potential cases visiting the website would be informed about the study, and a small proportion of them would volunteer to participate. The advertising efforts would have to last until a sufficient number of patients is enrolled in the study. The source population for this study would include all subjects that, if diagnosed with celiac disease, would search information on the Internet. Then controls could be recruited by posting information about the study in a number of disease-specific websites, say the websites of the Italian Associations for asthma, chronic bowel disease, and type 1 diabetes. Obviously, these diseases would have to be unrelated with the exposure of interest. Again, control participants would be self-selected volunteers, and the participation proportion would be expected to be very low. The association estimates obtained from an Internet-based case-control study of this type would be valid had the determinants of self-selection been similar among cases and controls or had these determinants been unrelated with the exposure status. Unfortunately, these conditions are both rather strong and difficult to check in the data.

A second approach for an Internet-based case-control study would begin with the definition of the source population. For the hypothetical Internet-based study on celiac disease determinants in Italy, a broad definition of the source population would include all Italian Internet users. Recruitment of case patients would then involve the website of the Italian Celiac Association, any other website of potential interest for celiac disease patients, as well as media campaigns. Case participants would again be strongly selected. Controls would have to be selected from the same source population, namely, from Internet users. Recruitment would thus involve media campaigns and links posted in various websites, not necessarily of interest for celiac patients.

A stricter definition of the source population could improve this study design. For example, the source population could include users of a specific website, say the website of a national newspaper. Everybody accessing the website would be invited to participate in a study, and case patients would include volunteers with celiac disease while control subjects would include all the other volunteers. It could be possible also to adopt a two-stage approach in which, in the first stage, a generic health Internet-based survey is offered to persons accessing the website and, at the second stage, cases with celiac disease and a sample of controls without celiac disease are further interviewed.

Irrespectively of whether a broader or a stricter definition for the source population is used, case and controls would again be strongly selected, and the critical issue would be whether the determinants of the selection are associated with the exposure of interest and if they differ between cases and controls.

This brief description of hypothetical Internet-based case-control studies emphasizes their vulnerability to selection bias. Careful methodological work and empirical testing is still needed before an Internet-based case-control study can actually be conducted. Apart from selection bias, however, Internet-based case-control studies would have other important limitations. Firstly, recruitment of incident cases through the Internet seems even more problematic and most likely an Internet-based case-control study would involve prevalent cases, with the well-known corresponding limitations. Secondly, even if a study manages to involve cases and controls from the same source population without selection bias, the use of the Internet introduces selection among cases. It is hard to predict if case participants would have a more or less serious disease, but the selection would most likely introduce problems of generalizability. Furthermore, it would be difficult to distinguish between factors causing the disease and factors affecting its severity or the patients' overall health status.

A case-crossover design (see chapter ▶Modern Epidemiological Study Designs of this handbook) which does not involve the selection of control subjects could be a more sensible option for an Internet-based study. An Internet-based case-crossover study on gout (Online Gout Study 2011) has indeed been launched in 2003 (Zhang et al. 2007). The existence of this study was advertised using Google links. Gout patients were invited to register and asked to complete a control-period questionnaire every 3 months investigating risk factors for gout attack in the preceding 2 days. Moreover, they were asked to complete an "attack questionnaire" if they were experiencing a gout attack. Participants were also provided a hard copy of the attack questionnaire in case they could not access the Internet during the gout attack, but in fact this option has been rarely used during the study.

As any other case-crossover study, this study aimed at evaluating trigger factors and acute effects (Hunter et al. 2006; Zhang et al. 2006). Exposure information was collected prospectively, because patients were recruited before having the actual gout attack, thus limiting the possibility of selection bias. It is also possible to imagine a different design, in which patients participate when they have an event, say a gout attack, and complete the exposure information for the control period/s retrospectively. Under this scenario, there is possibility of selection bias if their likelihood of accessing the study website and participating in the study depends on some of the trigger factors under investigation. For example, a patient could experience a gout attack after having used a diuretic. He or she might suspect that the diuretic was the cause of the gout attack and use the Internet to check for this hypothesis. Then, he or she accesses the study website, in which there might also be some general information about the disease, and decide to participate. Obviously, this would induce selection bias.

12.4.4 Intervention Studies

For a number of reasons, Internet recruitment fits very well with intervention studies, namely, (1) randomization is likely to cancel bias due to self-selection, (2) the randomization can be easily managed centrally through the website, also stratifying for a number of variables and patients' characteristics, and (3) pragmatic and explanatory trials become very similar in design and conduct. Indeed, Internet recruitment is the most correct setting in which to test tailored interventions offered via the Internet to unselected patients/populations. These types of interventions are becoming more and more frequent. We have carried out an admittedly cursory PubMed search restricted to clinical trials using "Internet-based" and "intervention" as the keywords. It revealed a clear trend in increasing number of papers with time (Fig. 12.9).

Internet-based trials in medicine have been conducted starting from the end of the 1990s and/or beginning of 2000s. In 2000, for example, McAlidon and colleagues started online recruitment for a feasibility study of an Internet-based clinical trial of glucosamine vs. placebo in patients with osteoarthritis of the knee (McAlindon et al. 2003). The main outcome was knee pain, assessed using a validated online self-completion questionnaire 12 weeks after intervention. Volunteers completed an eligibility screening questionnaire, and, if applicable, they were asked to send a signed hard copy of the informed consent and copies of medical records. Upon confirmation of the osteoarthritis of the knee, they were randomized into the treatment or placebo group. To those subjects included in the treatment group, the pills were mailed every second week. Although this study involved Internet-based recruitment, the intervention to be tested was not an online intervention (as it involved mailed pills) and relied on having access to the actual medical records to confirm the diagnosis and obtain detailed information on the disease. These aspects may decrease the possible methodological advantages of using the Internet.

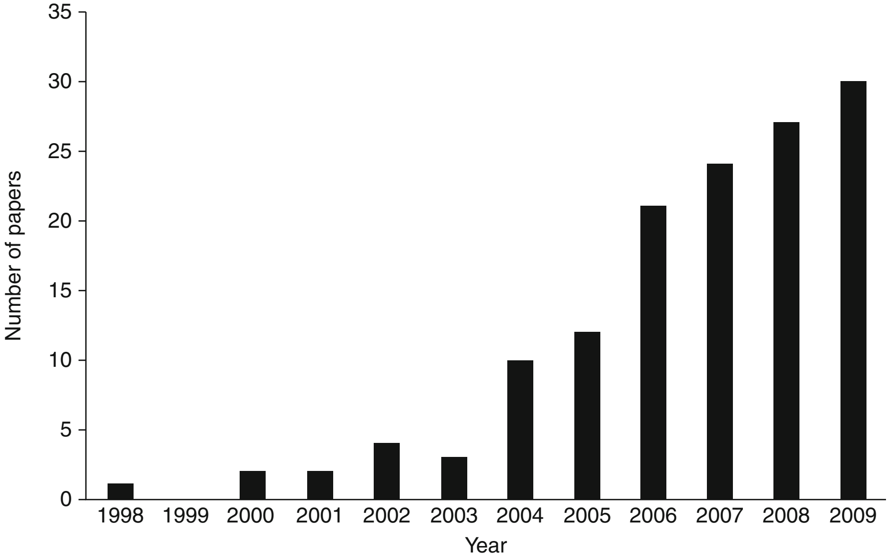


Fig. 12.9 Number of papers identified in PubMed using "Internet-based" and "intervention" as the keywords and restricting the search to clinical trials (search carried out in November 2010)

Studies testing online primary care interventions for health risk behaviors are more common, including those on diet and nutrition, physical activity, smoking habit, and alcohol consumption (Civljak et al. 2010; Portnoy et al. 2008; Vandelanotte et al. 2007; Webb et al. 2010). A recent Cochrane review on Internet-based interventions for smoking cessation identified 20 studies published until June 2010 (Civljak et al. 2010). Most of these studies used an Internet-based recruitment and all of them were published after 2004. Some of the trials compared an Internet intervention with either no intervention at all or an offline intervention; others compared different types of Internet interventions. In general, the study designs were very heterogeneous, and many studies had a relatively high proportion of dropouts. Authors concluded that the evidence of long-term benefits for programs delivered only by the Internet (as compared with offline interventions) is very limited, while there is some evidence that tailored (i.e., interventions specifically designed to meet the characteristics of a target individual or group) Internet interventions are more effective than non-tailored Internet interventions.

Rabius and colleagues advertised an intervention trial targeting smokers who wanted to quit smoking on the website of the American Cancer Society (Rabius et al. 2008). Potential participants were asked to complete a baseline questionnaire and provide informed consent. If eligible, they were randomized to receive access to one of five tailored interactive websites providing interventions for smoking cessation or a more static page set up at the American Cancer Society website serving as control treatment. The main outcome (successful abstinence for the last 30 days) was assessed 13 months after randomization, first e-mailing a survey questionnaire and then contacting by phone those who did not answer. Out of almost 6,500 individuals enrolled in the study, only 38% answered the follow-up questionnaire. Analyses were first conducted among the respondents only and then reconducted assuming that non-respondents did not quit smoking: in no case there was a significant difference in smoking cessation across the intervention arms. The high dropout rate between follow-ups, however, limited the study power.

Low completeness of follow-up was also observed in a trial set up on the Stop-tabac.ch website (Stop-tabac.ch 2011) to compare an original and a modified version of an online tailored program for smoking cessation (Etter 2005). The main outcome (smoking abstinence in the last 7 days) was assessed via e-mail 11 weeks after randomization, using up to three reminders. Almost 12,000 subjects were randomized in the study, with a response proportion at the follow-up questionnaire of 35%. In the analyses, non-respondents were assumed to still smoke. The original version of the program was found to be more effective, with 1 additional quitter every 26 participants.

Although many Internet-based randomized intervention studies have a high proportion of loss to follow-up, we do not think that this is a characteristic inherent to the study design. It is possible to increase participation by obtaining more detailed contact information at baseline (telephone number, cell phone number, address, second e-mail) and by recruiting a more committed and motivated population, for example, by having 1–2 follow-up questionnaires before randomization. Furthermore, in a pragmatic setting, a certain proportion of loss to follow-up should be expected.

Randomized trials recruiting participants through the Internet are feasible and, thanks to randomization, do not suffer from selection bias more than if a traditional approach of recruitment is used. Generalizability problems, namely, the restriction to self-selected Internet users, are minor when the intervention involves online programs. There are, however, some limitations. Firstly, it is difficult to have direct contact with participants and to include medical exams in the protocol, thus limiting the use of this study design in clinical settings. Secondly, in many studies, attrition to follow-up questionnaires was low and thus there is a need of methodological improvements to limit dropouts and increase motivation of the participants. Thirdly, Internet-based recruitment may lose its efficiency when the aim of the trial is to test an offline intervention.

12.4.5 Surveillance

The main aim of surveillance is to monitor trends in the rate of disease occurrence, both to gain insight in the current situation of an established disease and to detect outbreaks of emerging diseases.

In this paragraph, we will not go into detail in describing the aspects of surveillance, already illustrated in (see chapter ▶Emergency and Disaster Health Surveillance of this handbook). Instead, we will concentrate on surveillance conducted by means of the Internet as a way to collect data from sources not easily accessible by traditional surveillance.

The widespread diffusion of computers and of the Internet has provided a tool capable of the earliest possible detection of epidemics, giving allowance to a timely and maximally effective public health response worldwide. Surveillance data, as well as behavioral data, social contacts, and risk perception can be collected by exploiting new information and communication technology (ICT) techniques and methodologies to better understand the spread of infectious diseases (e.g., by collecting information on behavioral data to understand human immunodeficiency virus transmission) and obtain real-time data, which are crucial to rapidly identify public health emergencies, understand global trends, feed realistic data-driven models to assess the impact on the population, and optimize the allocation of resources.

Existing traditional disease surveillance systems have limitations. For example, in the case of influenza-like illness (ILI), monitoring methods rely on sentinel networks of physicians, laboratory scientists, public health professionals, and epidemiologists. Although they may mirror influenza activity, they cannot be implemented as real-time surveillance tools: either they only record proxy measures of influenza, or they contain unavoidable time delays between incidence and reporting. Traditional schemes require individuals to access health services and rely on the propensity of individuals to consult. Age-stratified rates of physician consultation may vary widely with different healthcare and health insurance systems. For non-severe diseases especially, only a minor (and unknown) fraction of all infected individuals sees a doctor, and frequently after a considerable delay, when a complication has occurred or in case a doctor's certificate is required. A web-based platform can be used to detect cases from those individuals who are less prone to consult a doctor when sick but agree to fill in a brief web survey about their symptoms. Moreover, traditional monitoring schemes typically lack uniform standards for clinical definitions that vary considerably between countries and even between reporters (EISN – European Influenza Surveillance Network). By using web surveys, standard platforms can be used across different countries to collect uniform data without major economic efforts.

In the following, we will focus on the case of influenza-like illnesses (ILI) for which Internet-based technologies have been successfully used. These Internet surveillance systems for ILI have been implemented in Belgium and the Netherlands in 2003 (under the name "Der Grote Griepmeting" – the Great Influenza Survey), in Portugal in 2005 ("Gripenet"), in Italy in 2007 ("Influweb"), in the UK in 2009 ("Flusurvey"), and in Sweden in 2010 ("Influensakoll") with the aim of measuring influenza activity and collecting important public health information in real time (e.g., during the 2009 H1N1 pandemic, the web platform detected the peak activity of the influenza more than 1 week in advance with respect to the general practitioners (GPs)). Platforms are now grouped under the umbrella of Influenzanet, forming a network of platforms to measure ILI in the community at a European level.

For each platform, registration of participants takes place through the web page (see flowchart in Fig. 12.10). Upon registration, following provision with a password-protected account, participants are asked to complete a baseline

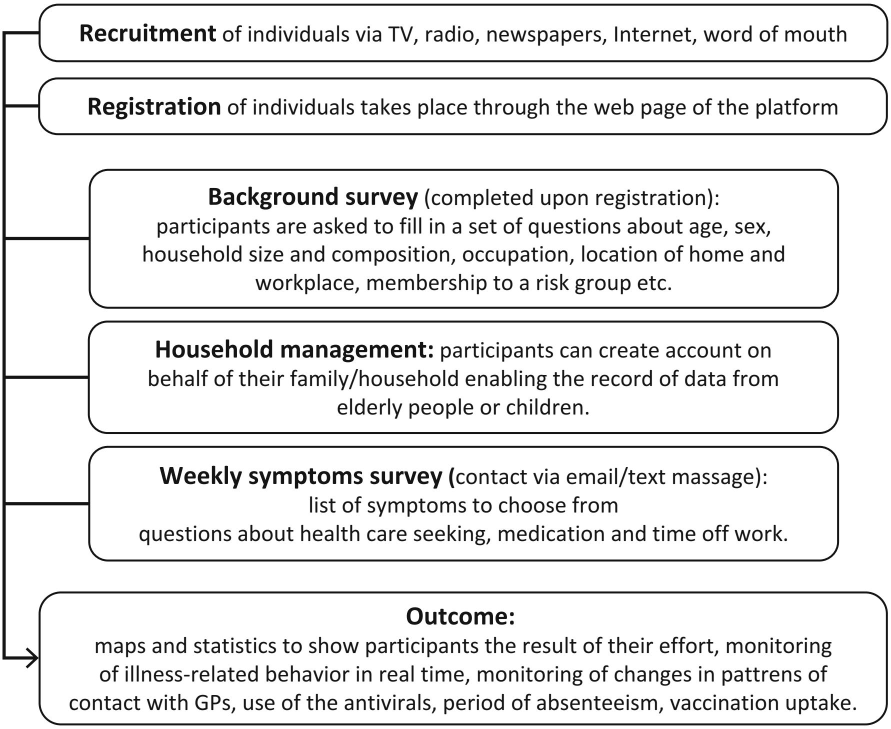

Recruitment of individuals via TV, radio, newspapers, Internet, word of mouth
Registration of individuals takes place through the web page of the platform
Background survey (completed upon registration): participants are asked to fill in a set of questions about age, sex, household size and composition, occupation, location of home and workplace, membership to a risk group etc.
Household management: participants can create account on behalf of their family/household enabling the record of data from elderly people or children.
Weekly symptoms survey (contact via email/text massage): list of symptoms to choose from questions about health care seeking, medication and time off work.
Outcome: maps and statistics to show participants the result of their effort, monitoring of illness-related behavior in real time, monitoring of changes in pattrens of contact with GPs, use of the antivirals, period of absenteeism, vaccination uptake.

Fig. 12.10 Diagram of recruitment and follow-up in the Influweb study

questionnaire with questions about age, sex, household size and composition, occupation, location of home and workplace, membership of a risk group, etc. Participants are also able to create accounts on behalf of their family/household enabling the entry of data from elderly people or children. Each week, participants are asked to complete a symptoms questionnaire as to whether or not they had symptoms in the previous week. To maintain participants' interest and remind them to complete the questionnaire, a newsletter containing influenza facts and news is sent each week. The platforms, called Internet monitoring systems (IMS), are updated on a daily basis during the whole influenza season with items including the estimated incidence and the spatial distribution of cases.

Results from the Belgian, Dutch, and Portuguese surveys have been analyzed under the name "Great Influenza Survey (GIS)" (Friesema et al. 2009). The estimated seasonal influenza incidence curves given by these systems were highly correlated with those obtained through the traditional surveillance method. These analyses also offer the encouraging indication that the Internet-based approaches can detect increased influenza activity more rapidly than surveillance-based reports by general practitioners (Marquet et al. 2006). Moreover, it is possible to estimate delays between the onset and the consultation dates and to detect changes in contact patterns and general behavior (Friesema et al. 2009).

During the 2009 H1N1 pandemic, IMSs have proved to be valuable tools in gaining an insight into the evolution of the pandemic in real time. In particular, the web-based system was launched in the UK during the first pandemic wave and went on collecting data until the end of the pandemic. Participants answering the symptoms questionnaires during the pandemic were asked more accurate follow-up questions about healthcare-seeking behavior, the delay of consultation with respect to the onset of symptoms; if they took time off work and for how long; if they took antiviral medication; if they were willing to be vaccinated against H1N1; etc. (Tilston et al. 2010).

Since participants would select their symptoms from a prespecified list, it was also possible to test different definitions of ILI, and to compare the resulting incidence with the one estimated by the Health Protection Agency (HPA). During the pandemic, to get a clearer picture of the epidemic evolution, the HPA asked for random testing of patients accessing different healthcare settings, which allowed evaluation of the true number of cases and thus adjustment of the estimates, by means of a method that was expensive and induced further delays in the data stream. The IMS to monitor ILI in the community was a direct and timely alternative, providing an incidence curve with a timing of the peak being close to the adjusted HPA case estimates. In making comparisons between web-based system estimates and HPA case numbers, results suggest that trends can be captured by the IMS even more reliably than standard GP-based systems, even though it remains unclear how accurate they are for estimating the absolute level of incidence (Tilston et al. 2010).

The UK web-based surveillance platform that ran continuously from July 2009 to March 2010 (i.e., during both the first (summer) and the second (autumn) pandemic wave in England in 2009) has also detected changes in healthcare-seeking behavior between the two waves (Brooks-Pollock et al. 2011). These behavioral modifications, due to changing scientific information, media coverage, and public anxiety, affected official case estimates. The web-based platform was able to detect a decrease from 43% to 32% in the percentage of individuals with ILI symptoms who sought medical attention from start to the end of the epidemic. Adjusting official numbers accordingly, it was possible to estimate that there were 1.1 million symptomatic cases in England, over 40% more cases than previously estimated, and that the autumn epidemic wave was 45% bigger than previously thought.

A further aspect for which surveillance is crucial is the development of accurate prediction models. When an outbreak occurs, usually short- and long-term predictions are based on observed data (provided by GP consultations or death/hospitalization records) combined with mathematical models updated as more data arise. For example, Baguelin et al. (2010) carried out a real-time assessment of the effectiveness and cost-effectiveness of alternative influenza A/H1N1v vaccination strategies. To generate plausible autumn scenarios under different vaccination options, they fitted a transmission dynamic model using the estimated number of cases determined by means of a web-based surveillance platform, calls to the UK National Pandemic Flu Service, and GP calls and consultations. In this specific case, data collected by means of the web platform were used to estimate the proportion of ILI cases calling the GP during the influenza A/H1N1v in the UK.

One of the important limitations that remain related to Internet-based surveillance is the lack of medical or laboratory confirmation of diagnosis. Data collected by means of a web platform will never be able to replace the virological analysis or clinical diagnosis carried out by GP surveillance.

In direct contact with the patient, the GP can exclude other diseases than ILI, and the virological analysis can give further information about the subtype of influenza virus. A possibility to overcome the latter limitation could be to send self-sampling kits to a selected subset of the Internet-based system participants. An attempt in this direction has already been made (Cooper et al. 2008), and this possibility should be explored further. Another possible limitation of Internet-based surveillance is that participants are not representative of the general population. This issue has been addressed by reweighting the sample according to the age and sex distribution in the general population (Tilston et al. 2010), although it is never possible to exclude that participation in an Internet-based system is positively or negatively associated with the risk of ILI.

In conclusion, Internet-based surveillance has the potential to capture a wider range of ILI cases than traditional surveillance, as well to track changes in healthcare attendance patterns in real time. Even though Internet-based surveillance has limitations and cannot replace traditional GP-based surveillance, it can provide an important support to enable the collection of valuable additional information, both in ordinary surveillance and during public health crises when the sentinel GPs surveillance and public health systems are under stress. While ILI has been used in the early deployment of the system, in subsequent years, these IMSs will consider other diseases and infections.

12.5 Web 2.0

Very recently, the Internet has started offering new possibilities for epidemiological research based on the so-called Web 2.0, which refers to the active generation of contents by the Internet users through various means including online communities, web searches, social networks, etc. (Lee 2010). For example, these new communication and information habits over the Internet may be used to quantify outbreaks of specific diseases (Eysenbach 2002, 2009). A well-known example is the use of Google searches to obtain real-time estimates of influenza-like illnesses (ILI) in the United States (Ginsberg et al. 2009): Ginsberg and colleagues developed a method to identify automatically ILI-related web search queries and use them to estimate ILI weakly percentages; these estimates had a correlation above 0.90 with data obtained from the US Surveillance Network of the CDC (Centre for Disease Control and Prevention).

The potential sources of information available in the Web 2.0 are growing fast. As examples, data on disease outbreaks can be obtained from chat rooms, blogs, press release, or Facebook (Brownstein et al. 2009, 2010; Eysenbach 2009), while exposure data, say on air pollution, can be obtained continuously from sensors worn by self-selected volunteers, and automatically transferred in real-time for model analyses (Mobile Environmental Sensing System Across Grid Environments, Message 2008; Mobile Air Quality Monitoring Network, Institute for Software Integrated Systems 2011). There are, however, obvious limitations in using these sources of data, including issues of generalizability, bias, and exposure and outcome measurement. Currently, areas of epidemiological research which may actually benefit from the use of Web 2.0 remain to be explored and identified.

12.6 Conclusions

Fifteen years ago, the Internet was advocated as a promising tool for epidemiological research (Rothman et al. 1997). Since then, some Internet-based studies have been conducted, including a number of surveys and intervention trials along with few cohort studies. Nevertheless, the use of the Internet in epidemiological research is still very limited. Van Gelder and colleagues have recently reviewed analytical epidemiological papers published in 2008–2009 in four top general medical journals (the British Medical Journal, the Journal of the American Medical Association, The Lancet, and the New England Journal of Medicine) and in three top epidemiological journals (American Journal of Epidemiology, Epidemiology, and International Journal of Epidemiology), finding that only about 1% of the scrutinized 2,094 studies used Internet-based questionnaires (van Gelder et al. 2010). Since online questionnaires can be also used in studies which do not recruit online, the actual number of Internet-based studies is smaller. For this chapter, we have scrutinized papers published between September 21, 2009, and September 20, 2010, in 11 leading epidemiological journals (Environmental Health Perspective, American Journal of Epidemiology, Epidemiology, International Journal of Epidemiology, American Journal of Public Health, American Journal of Preventive Medicine, European Journal of Epidemiology, Preventive Medicine, Journal of Epidemiology and Community Health, Journal of Clinical Epidemiology, Annals of Epidemiology). We have identified only two papers, both methodological, concerning an Internet-based cohort (Huybrechts et al. 2010) and an Internet-based survey (Klovning et al. 2009).

These data demonstrate that the use of the Internet has not yet routinely entered the epidemiological practice. Moreover, most publications of Internet-based studies deal with feasibility or proof-of-concept studies.

Evidence on the validity and reliability of web-based questionnaires has started accumulating (Apovian et al. 2010; Brigham et al. 2009; Donker et al. 2009; Miller et al. 2002; Rankin et al. 2008; Touvier et al. 2010, 2011; West et al. 2006), and some studies have successfully used mixed methods involving both web-based and mailed questionnaires for the follow-up of traditionally recruited cohorts (Ekman et al. 2006; Russell et al. 2010). However, methodological research on the use of the Internet to recruit unspecified populations in the context of the various types of study designs is still in its infancy. In detail, we need to better understand when an Internet-based survey can be carried out and if it is possible to tackle some of its inherent problems of selection bias; we need to further investigate the impact of baseline selection in Internet-based cohort studies, as well as evaluate the determinants of completeness of follow-up; we should refine our ability to advertise the existence of Internet-based studies to the relevant source population; we should develop methods to improve attrition in Internet-based randomized studies; we should understand if it is possible to conduct Internet-based case-control studies; and we should understand when surveillance can (or should) be carried out via the Internet.

In our opinion the short-, medium-term agenda for Internet-based research applied to epidemiology should prioritize innovative field work and methodological studies on the acquired data. In other words, the debate on whether Internet-based research is valid or not should overcome a-priori formed opinions and become more evidence-based where this chapter has indicated that randomized trials, cohort studies, and surveillance may be successfully carried out using the Internet.

Acknowledgements This work was partially supported by the Compagnia San Paolo and the Piedmont Region. We thank Prof. Neil Pearce for valuable comments and Jose' Lobo for valuable help.

References

Andersson G, Lindvall N, Hursti T, Carlbring P (2002) Hypersensitivity to sound (hyperacusis): a prevalence study conducted via the internet and post. Int J Audiol 41:545–554

Apovian CM, Murphy MC, Cullum-Dugan D, Lin PH, Gilbert KM, Coffman G, Jenkins M, Bakun P, Tucker KL, Moore TJ (2010) Validation of a web-based dietary questionnaire designed for the DASH (dietary approaches to stop hypertension) diet: the DASH online questionnaire. Public Health Nutr 13:615–622

Baehring TU, Schulze H, Bornstein SR, Scherbaum WA (1997) Using the World Wide Web – a new approach to risk identification of diabetes mellitus. Int J Med Inform 46:31–39

Baer A, Saroiu S, Koutsky LA (2002) Obtaining sensitive data through the Web: an example of design and methods. Epidemiology 13:640–645

Baguelin M, Hoek AJ, Jit M, Flasche S, White PJ, Edmunds WJ (2010) Vaccination against pandemic influenza A/H1N1v in England: a real-time economic evaluation. Vaccine 28:2370–2384

Berrens RP, Jenkins-Smith H, Silva C, Weimer DL (2003) The advent of internet surveys for political research: a comparison of telephone and internet samples. Political Anal 11:1–22

Bowen AM, Williams ML, Daniel CM, Clayton S (2008) Internet based HIV prevention research targeting rural MSM: feasibility, acceptability, and preliminary efficacy. J Behav Med 31: 463–477

Brigham J, Lessov-Schlaggar CN, Javitz HS, Krasnow RE, McElroy M, Swan GE (2009) Test-retest reliability of web-based retrospective self-report of tobacco exposure and risk. J Med Internet Res 11:e35

Brooks-Pollock E, Tilston N, Edmunds WJ, Eames KT (2011) Using an online survey of healthcare-seeking behaviour to estimate the magnitude and severity of the 2009 H1N1v influenza epidemic in England. BMC Infect Dis 11:68

Brownstein JS, Freifeld CC, Madoff LC (2009) Digital disease detection–harnessing the Web for public health surveillance. N Engl J Med 360:2153–2155, 2157

Brownstein JS, Freifeld CC, Chan EH, Keller M, Sonricker AL, Mekaru SR, Buckeridge DL (2010) Information technology and global surveillance of cases of 2009 H1N1 influenza. N Engl J Med 362:1731–1735

Buchanan T, Smith JL (1999) Using the internet for psychological research: personality testing on the World Wide Web. Br J Psychol 90(Pt 1):125–144

Civljak M, Sheikh A, Stead LF, Car J (2010) Internet-based interventions for smoking cessation. Cochrane Database Syst Rev:CD007078

Cooper DL, Smith GE, Chinemana F, Joseph C, Loveridge P, Sebastionpillai P, Gerard E, Zambon M (2008) Linking syndromic surveillance with virological self-sampling. Epidemiol Infect 136:222–224

Craiglist (2011) www.craiglist.com. Accessed 9 May 2011

Dillman DA, Smyth JD (2007) Design effects in the transition to web-based surveys. Am J Prev Med 32:S90–S96

Dillman DA, Smyth JD, Christian LM (2009) Internet, mail, and mixed-mode surveys. The taiolred design method. Wiley, Haboken

Donker T, van Straten A, Marks I, Cuijpers P (2009) A brief Web-based screening questionnaire for common mental disorders: development and validation. J Med Internet Res 11:e19

Ekman A, Dickman PW, Klint A, Weiderpass E, Litton JE (2006) Feasibility of using web-based questionnaires in large population-based epidemiological studies. Eur J Epidemiol 21:103–111

Ekman A, Klint A, Dickman PW, Adami HO, Litton JE (2007) Optimizing the design of web-based questionnaires – experience from a population-based study among 50,000 women. Eur J Epidemiol 22:293–300

Elford J, Jeannin A, Spencer B, Gervasoni JP, van de Laar MJ, Dubois-Arber F (2009) HIV and STI behavioural surveillance among men who have sex with men in Europe. Euro Surveill 14:pii:19414

Etter JF (2005) Comparing the efficacy of two internet-based, computer-tailored smoking cessation programs: a randomized trial. J Med Internet Res 7:e2

Etter JF, Perneger TV (2001) A comparison of cigarette smokers recruited through the internet or by mail. Int J Epidemiol 30:521–525

Etter JF, Neidhart E, Bertrand S, Malafosse A, Bertrand D (2005) Collecting saliva by mail for genetic and cotinine analyses in participants recruited through the internet. Eur J Epidemiol 20:833–838

Etter JF, Hoda JC, Perroud N, Munafo M, Buresi C, Duret C, Neidhart E, Malafosse A, Bertrand D (2009) Association of genes coding for the alpha-4, alpha-5, beta-2 and beta-3 subunits of nicotinic receptors with cigarette smoking and nicotine dependence. Addict Behav 34:772–775

Evans AR, Wiggins RD, Mercer CH, Bolding GJ, Elford J (2007) Men who have sex with men in Great Britain: comparison of a self-selected internet sample with a national probability sample. Sex Transm Infect 83:200–205; discussion 205

Eysenbach G (2002) Infodemiology: the epidemiology of (mis)information. Am J Med 113:763–765

Eysenbach G (2009) Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. J Med Internet Res 11:e11

Friesema IH, Koppeschaar CE, Donker GA, Dijkstra F, van Noort SP, Smallenburg R, van der Hoek W, van der Sande MA (2009) Internet-based monitoring of influenza-like illness in the general population: experience of five influenza seasons in The Netherlands. Vaccine 27:6353–6357

gaydar (2011) http://www.gaydar.co.uk/. Accessed 9 May 2011

gay.com (2011) http://www.gay.com/. Accessed 9 May 2011

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009) Detecting influenza epidemics using search engine query data. Nature 457:1012–1014

Glymour MM (2006) Using causal diagrams to understand common problems in social epidemiology. In: Oakes JM, Kaufman JS (eds) Methods in social epidemiology. Jossey-Bass, San Francisco, pp 387–422

Gosling SD, Vazire S, Srivastava S, John OP (2004) Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. Am Psychol 59:93–104

Greenland S (2003) Quantifying biases in causal models: classical confounding vs collider-stratification bias. Epidemiology 14:300–306

Hercberg S, Castetbon K, Czernichow S, Malon A, Mejean C, Kesse E, Touvier M, Galan P (2010) The Nutrinet-Sante Study: a web-based prospective study on the relationship between nutrition and health and determinants of dietary patterns and nutritional status. BMC Public Health 10:242

Hernán MA, Hernández-Diaz S, Robins JM (2004) A structural approach to selection bias. Epidemiology 15:615–625

Hilsden RJ, Meddings JB, Verhoef MJ (1999) Complementary and alternative medicine use by patients with inflammatory bowel disease: An internet survey. Can J Gastroenterol 13:327–332

Hirshfield S, Remien RH, Humberstone M, Walavalkar I, Chiasson MA (2004) Substance use and high-risk sex among men who have sex with men: a national online study in the USA. AIDS Care 16:1036–1047

Hunter DJ, York M, Chaisson CE, Woods R, Niu J, Zhang Y (2006) Recent diuretic use and the risk of recurrent gout attacks: the online case-crossover gout study. J Rheumatol 33:1341–1345

Huybrechts KF, Mikkelsen EM, Christensen T, Riis AH, Hatch EE, Wise LA, Sorensen HT, Rothman KJ (2010) A successful implementation of e-epidemiology: the Danish pregnancy planning study ‘Snart-Gravid’. Eur J Epidemiol 25:297–304

Influenzanet (2011) www.influweb.it. Accessed 9 May 2011

Institute for Software Integrated Systems (2011) Mobile Air Quality Monitoring Network. http://www.isis.vanderbilt.edu/projects/maqumon. Accessed 9 May 2011

Internet World Stats (2011) http://www.internetworldstats.com. Accessed 9 May 2011

Kelly MA, Oldham J (1997) The internet and randomised controlled trials. Int J Med Inform 47: 91–99

Klovning A, Sandvik H, Hunskaar S (2009) Web-based survey attracted age-biased sample with more severe illness than paper-based survey. J Clin Epidemiol 62:1068–1074

Kongsved SM, Basnov M, Holm-Christensen K, Hjollund NH (2007) Response rate and completeness of questionnaires: a randomized study of internet versus paper-and-pencil versions. J Med Internet Res 9:e25

Kraut R, Olson J, Banaji M, Bruckman A, Cohen J, Couper M (2004) Psychological research online: report of Board of Scientific Affairs’ Advisory Group on the Conduct of Research on the Internet. Am Psychol 59:105–117

Kushi LH, Finnegan J, Martinson B, Rightmyer J, Vachon C, Yochum L (1997) Epidemiology and the internet. Epidemiology 8:689–690

Lee BK (2010) Epidemiologic research and Web 2.0 – the user-driven Web. Epidemiology 21: 760–763

Marcus U, Schmidt AJ, Hamouda O, Bochow M (2009) Estimating the regional distribution of men who have sex with men (MSM) based on internet surveys. BMC Public Health 9:180

Marquet RL, Bartelds AI, van Noort SP, Koppeschaar CE, Paget J, Schellevis FG, van der Zee J (2006) Internet-based monitoring of influenza-like illness (ILI) in the general population of the Netherlands during the 2003–2004 influenza season. BMC Public Health 6:242

McAlindon T, Formica M, Kabbara K, LaValley M, Lehmer M (2003) Conducting clinical trials over the internet: feasibility study. BMJ 327:484–487

Message (2008) Mobile Environmental Sensing System Across Grid Environments. http://bioinf.ncl.ac.uk/message/. Accessed 9 May 2011

Mikkelsen EM, Hatch EE, Wise LA, Rothman KJ, Riis A, Sorensen HT (2009) Cohort profile: the Danish Web-based Pregnancy Planning Study – ‘Snart-Gravid’. Int J Epidemiol 38:938–943

Miller ET, Neal DJ, Roberts LJ, Baer JS, Cressler SO, Metrik J, Marlatt GA (2002) Test-retest reliability of alcohol measures: is there a difference between internet-based assessment and traditional methods? Psychol Addict Behav 16:56–63

Miller PG, Johnston J, Dunn M, Fry CL, Degenhardt L (2010) Comparing probability and non-probability sampling methods in Ecstasy research: implications for the internet as a research tool. Subst Use Misuse 45:437–450

NINFEA Project (2011) www.progettoninfea.it. Accessed 9 May 2011

Online Gout Study (2011) https://dcc2.bumc.bu.edu/goutstudy/. Accessed 9 May 2011

Paolotti D, Giannini C, Colizza A, Vespignani A (2010) Internet-based monitoring system for Influenza-like illness: H1N1 surveillance in Italy. In: Proceedings of Ehealt, Casablanca

Pepine CJ, Handberg-Thurmond E, Marks RG, Conlon M, Cooper-DeHoff R, Volkers P, Zellig P (1998) Rationale and design of the International Verapamil SR/Trandolapril Study (INVEST): an internet-based randomized trial in coronary artery disease patients with hypertension. J Am Coll Cardiol 32:1228–1237

Pizzi C, De Stavola B, Merletti F, Bellocco R, Dos Santos Silva I, Pearce N, Richiardi L (2011) Sample selection and validity of exposure-disease association estimates in cohort studies. J Epidemiol Community Health 65:407–411

Pizzi C, De Stavola B, Pearce N, Lazzarato F, Ghiotti P, Merletti F, Richiardi L (2012) Selection bias and patterns of confounding in cohort studies: the case of the NINFEA web-based birth cohort. J Epidemiol Community Health 66:976–981

Portnoy DB, Scott-Sheldon LA, Johnson BT, Carey MP (2008) Computer-delivered interventions for health promotion and behavioral risk reduction: a meta-analysis of 75 randomized controlled trials, 1988–2007. Prev Med 47:3–16

Rabius V, Pike KJ, Wiatrek D, McAlister AL (2008) Comparing internet assistance for smoking cessation: 13-month follow-up of a six-arm randomized controlled trial. J Med Internet Res 10:e45

Ramo DE, Hall SM, Prochaska JJ (2010) Reaching young adult smokers through the internet: comparison of three recruitment mechanisms. Nicotine Tob Res 12:768–775

Rankin KM, Rauscher GH, McCarthy B, Erdal S, Lada P, Il'yasova D, Davis F (2008) Comparing the reliability of responses to telephone-administered versus self-administered Web-based surveys in a case-control study of adult malignant brain cancer. Cancer Epidemiol Biomarkers Prev 17:2639–2646

Rhodes SD, Bowie DA, Hergenrather KC (2003) Collecting behavioural data using the world wide web: considerations for researchers. J Epidemiol Community Health 57:68–73

Richiardi L, Baussano I, Vizzini L, Douwes J, Pearce N, Merletti F (2007) Feasibility of recruiting a birth cohort through the internet: the experience of the NINFEA cohort. Eur J Epidemiol 22:831–837

Ross MW, Mansson SA, Daneback K, Cooper A, Tikkanen R (2005) Biases in internet sexual health samples: comparison of an internet sexuality survey and a national sexual health survey in Sweden. Soc Sci Med 61:245–252

Rosser BR, Oakes JM, Horvath KJ, Konstan JA, Danilenko GP, Peterson JL (2009) HIV sexual risk behavior by men who use the internet to seek sex with men: results of the Men's INTernet Sex Study-II (MINTS-II). AIDS Behav 13:488–498

Rothman KJ, Cann CI, Walker AM (1997) Epidemiology and the internet. Epidemiology 8: 123–125

Ruddy KJ, Gelber S, Shin J, Garber JE, Rosenberg R, Przypysny M, Partridge AH (2010) Genetic testing in young women with breast cancer: results from a Web-based survey. Ann Oncol 21:741–747

Russell CW, Boggs DA, Palmer JR, Rosenberg L (2010) Use of a web-based questionnaire in the Black Women's Health Study. Am J Epidemiol 172:1286–1291

Sanders AR, Levinson DF, Duan J, Dennis JM, Li R, Kendler KS, Rice JP, Shi J, Mowry BJ, Amin F, Silverman JM, Buccola NG, Byerley WF, Black DW, Freedman R, Cloninger CR, Gejman PV (2010) The internet-based MGS2 control sample: self report of mental illness. Am J Psychiatry 167:854–865

Schleyer TK, Forrest JL (2000) Methods for the design and administration of web-based surveys. J Am Med Inform Assoc 7:416–425

Silver RC, Holman EA, McIntosh DN, Poulin M, Gil-Rivas V (2002) Nationwide longitudinal study of psychological responses to September 11. JAMA 288:1235–1244

Simmons RD, Ponsonby AL, van der Mei IA, Sheridan P (2004) What affects your MS? Responses to an anonymous, internet-based epidemiological survey. Mult Scler 10:202–211

Skitka LJ, Sargis EG (2006) The internet as psychological laboratory. Annu Rev Psychol 57: 529–555

Soetikno RM, Mrad R, Pao V, Lenert LA (1997) Quality-of-life research on the internet: feasibility and potential biases in patients with ulcerative colitis. J Am Med Inform Assoc 4:426–435

Stop-tabac.ch (2011) http://www.stop-tabac.ch/fra/. Accessed 9 May 2011

The Early Life Factors Study of Childhood Diseases (2011) http://www.elfs.org.nz/. Accessed 9 May 2011

Tilston NL, Eames KT, Paolotti D, Ealden T, Edmunds WJ (2010) Internet-based surveillance of Influenza-like-illness in the UK during the 2009 H1N1 influenza pandemic. BMC Public Health 10:650

Touvier M, Mejean C, Kesse-Guyot E, Pollet C, Malon A, Castetbon K, Hercberg S (2010) Comparison between web-based and paper versions of a self-administered anthropometric questionnaire. Eur J Epidemiol 25:287–296

Touvier M, Kesse-Guyot E, Mejean C, Pollet C, Malon A, Castetbon K, Hercberg S (2011) Comparison between an interactive web-based self-administered 24 h dietary record and an interview by a dietitian for large-scale epidemiological studies. Br J Nutr 105:1055–1064

Treadwell JR, Soetikno RM, Lenert LA (1999) Feasibility of quality-of-life research on the internet: a follow-up study. Qual Life Res 8:743–747

Turner C, Bain C, Schluter PJ, Yorkston E, Bogossian F, McClure R, Huntington A (2009) Cohort profile: The Nurses and Midwives e-Cohort Study–a novel electronic longitudinal study. Int J Epidemiol 38:53–60

United Nations Economic Commission for Europe (UNECE) (2011) http://w3.unece.org/pxweb/dialog/varval.asp?ma=02_GEICT_InternetUse_r&path=../DATABASE/STAT/30-GE/09-09-Science_ICT&lang=1. Accessed 9 May 2011

Vandelanotte C, Spathonis KM, Eakin EG, Owen N (2007) Website-delivered physical activity interventions a review of the literature. Am J Prev Med 33:54–64

van Gelder MM, Bretveld RW, Roeleveld N (2010) Web-based questionnaires: the future in epidemiology? Am J Epidemiol 172:1292–1298

Wang J, Etter JF (2004) Administering an effective health intervention for smoking cessation online: the international users of Stop-Tabac. Prev Med 39:962–968

Webb TL, Joseph J, Yardley L, Michie S (2010) Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 12:e4

West R, Gilsenan A, Coste F, Zhou X, Brouard R, Nonnemaker J, Curry SJ, Sullivan SD (2006) The ATTEMPT cohort: a multi-national longitudinal study of predictors, patterns and consequences of smoking cessation; introduction and evaluation of internet recruitment and data collection methods. Addiction 101:1352–1361

Zhang Y, Woods R, Chaisson CE, Neogi T, Niu J, McAlindon TE, Hunter D (2006) Alcohol consumption as a trigger of recurrent gout attacks. Am J Med 119:800.e813–800.e808

Zhang Y, Chaisson CE, McAlindon T, Woods R, Hunter DJ, Niu J, Neogi T, Felson DT (2007) The online case-crossover study is a novel approach to study triggers for recurrent disease flares. J Clin Epidemiol 60:50–55

Part II

Methodological Approaches in Epidemiology

Design and Planning of Epidemiological Studies

13

Wolfgang Ahrens, Iris Pigeot, and Pascal Wild

Contents

W. Ahrens
Department of Epidemiological Methods and Etiologic Research, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

Faculty of Mathematics/Computer Science, University of Bremen, Bremen, Germany

I. Pigeot (✉)
Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

Faculty of Mathematics/Computer Science, University of Bremen, Bremen, Germany

P. Wild
Direction scientifique (Scientific Management), Occupational Health and Safety Institute (INRS), Vandoeuvre-lès-Nancy, France

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_13,

13.1 Introduction

This chapter deals with basic principles and practical issues in designing and planning epidemiological studies. Although most of the practical issues are consequences of the theoretical principles of epidemiology as presented in this handbook as well as standard textbooks (Breslow and Day 1981, 1987; Gordis 2009; Rothman and Greenland 1998; Rothman et al. 2008; Miettinen 1985) and in different chapters (▸Basic Concepts, ▸Cohort Studies, ▸Case-Control Studies, ▸Confounding and Interaction, ▸Sample Size Determination in Epidemiological Studies, ▸Measurement Error, and ▸Missing Data) of this handbook, the emphasis here is on how to proceed practically when planning a study.

This chapter is based on our experience in conducting epidemiological studies and on a series of references in which many of the concerns in the practical planning of studies have been described at length. Among the key sources we used are the books by Hernberg (1992), White et al. (2008), Olsen et al. (2010), and Kreienbrock et al. (2012). The series of papers by Wacholder et al. (1992a, b, c) have also inspired much of our writing. It will start with a section on early planning in which the general setting of any study is described as well as the key planning document, that is, the study protocol. The second section is devoted to the choice and implementation of an actual design where we focus on cohort and case-control studies. The next section focuses on data collection and biological samples, both, with respect to the exposure and the disease outcome. The final section is devoted to practical issues and gives a list of topics arising in all studies that may not always get the appropriate attention while planning an epidemiological study.

13.2 Scientific Framework

13.2.1 Study Questions and Hypotheses

The first step in the planning of an epidemiological study is the definition of the problem, that is, the statement of the hypotheses and the justification of their relevance. Researchers must ensure that they have a clear view of the problem at the general level. At this conceptual level, a problem takes the form: "does X cause Y?" or "how much will a certain amount of exposure to X affect Y?" (Hernberg 1992, Chap. 4). To answer such questions, we first have to define the target population, say all male residents of the USA aged 18 to 69 years (see Fig. 13.1). For pragmatic reasons this population has to be narrowed down to the so-called study base, for example, all male residents of six selected US communities aged 18 to 69 years. A random sample drawn from this study base may serve as the study sample or study population. It should be stressed that in etiological as well as in descriptive studies, researchers are interested in the relationship between factors and disease which are valid beyond the study population, that is, which are informative for the study base and for the target population which is represented by the study base. Ultimately,

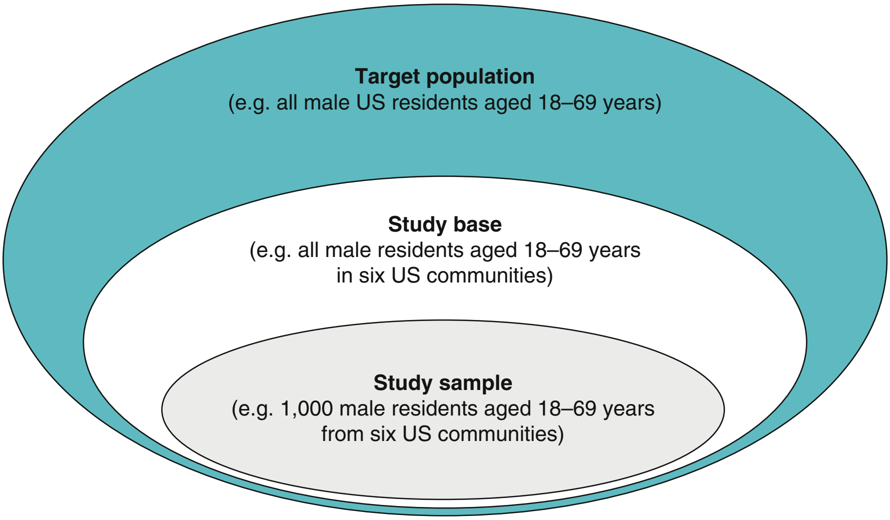

Target population
(e.g. all male US residents aged 18–69 years)
Study base
(e.g. all male residents aged 18–69 years
in six US communities)
Study sample
(e.g. 1,000 male residents aged 18–69 years
from six US communities)

Fig. 13.1 Hierarchy of target population, study base, and study sample

they are interested in the particular morbidity experience of the target population as far as it can be extrapolated to other populations.

The general interest of the investigator first has to be translated into precisely formulated, written objectives. A limited number of study objectives should be defined. These objectives may be of two kinds.

First, the study may be focused on specific questions with a predefined hypothesis. For instance, "does exposure to extremely low-frequency (ELF) electromagnetic fields cause childhood leukemia?" Here the hypothesis should be formulated as a series of operational questions. One can specify the ELF fields in a variety of ways both qualitative (yes/no or low/medium/high) and quantitative (e.g., present intensity, mean intensity over the last years, or cumulative exposure). Other predefined operational hypotheses may include subgroup analyses based on the disease subtype. These operational hypotheses will have to be decided by confirmatory tests. Additional findings beyond these primary hypotheses can then be clearly sorted out.

Second, the study may be designed to generate new hypotheses that will be investigated by exploratory analyses such as the following "what occupations are associated with an increased risk of laryngeal cancer?" Even in this case a predefined list of exploratory questions is useful. In the present example, this would consist in a list of occupations considered. Again an unsuspected excess in an occupation not considered a priori as entailing an excess risk should be reported as such.

13.2.2 Scientific Background

It is of crucial importance to undertake a thorough literature search and to know the literature in detail before planning any new project. Occasionally, the literature review may show that the study question has already been answered and that further data collection would be wasted.

Typically, the formulation of the study question (cf. Sect. 13.2.1) and the review of the scientific background are interacting steps in the initial phase of a new study that lead to the development of a study protocol (see Sect. 13.2.3). The scientific background has to support the research question and must justify its relevance in the light of the state of the art. This relates to both the subject matter to be investigated and the methodological approaches and instruments to be used.

Evaluating epidemiological evidence from the literature is often challenging even for experienced researchers. Because a single positive finding may be a chance finding, a complete literature search, including negative results, should be conducted. One should consider systematic errors (biases) and confounders that may have led to a particular result in previous studies. Several independent studies using the same design and the same procedure of data collection may have similar results due to common biases or confounding. It is therefore important that the sources of spurious results are identified and controlled in subsequent investigations. It would be ineffective to simply replicate previous studies, without consideration of new research questions raised by previous studies that could not be addressed because the information was not collected. A large body of literature on how to perform a systematic literature synthesis is available; see chapter ▶Meta-Analysis in Epidemiology of this handbook and references therein.

The literature review should, however, not be restricted to epidemiological studies, but has to encompass a range of topics from biological mechanisms or biomarkers related to the hypothesis under study to techniques of exposure measurement and other methodological aspects relevant to answer the study question(s).

13.2.3 The Study Protocol

An epidemiological study is generally a complex undertaking of long duration, requiring substantial personnel and financial resources. The success of a study depends on a careful preparation. It is self-evident that such an undertaking cannot be done without a written study protocol. The study protocol gives the scientific reasoning behind a study, specifies the research question, and describes the methods to be applied including those for statistical data analysis.

After describing the scientific rationale, the study background, and the specific objectives, the study protocol should define the study design including the precise study base and an estimate of the corresponding statistical power to achieve the objectives. Finally it should summarize the main tasks, their sequence, and the means to achieve the corresponding deliverables. This entails, for instance, procedures for identifying the outcome parameter, measuring of exposure and confounders, data management, and all steps taken for quality assurance. Further practical details of the fieldwork and the study logistics should be worked out separately in a manual of operations. The study protocol should also address the strategy for statistical analysis, quality control and training, time schedule, ethical considerations and principles of data protection, the general project organization, publication rules, and the financial breakdown.

According to Miettinen (1985), a study protocol serves five purposes, namely to

- crystallize the project to the researchers themselves
- give referees the possibility to review the project (especially for funding)
- inform and educate all those taking part in the project
- ensure the main researchers do not forget any details of the plan in the course of the study
- document the procedures of the project for the future.

It should be noted that Miettinen obviously did not differentiate between study protocol and manual of operations as to the authors' view the last three purposes are more appropriately addressed by a complementary manual of operations (see Sect. 13.7.4).

In summary, a protocol must be so detailed that an independent researcher could carry out the study based on it. An outline of the issues to be covered in the study protocol is given in Box 13.1. Most of them will be discussed in what follows. For further details we refer to chapter ▶Quality Control and Good Epidemiological Practice of this handbook.

Box 13.1. Overview of a study plan and corresponding key problems to be addressed in the planning phase

1. *Research questions and working hypotheses*
   - Relevance/previous knowledge
   - Choice of an appropriate target population
   - Precise definition of endpoint
   - Precise definition of independent variables and confounders
   - Choice of statistical measures (proportions, means, risks)
   - Translation into a hypothesis that can be tested statistically
   - Confirmatory testing of hypothesis/exploratory analyses

2. *Study design*
   - Optimal design (theoretical)
   - Practical limitations/feasibility
   - Justification of chosen design

3. *Target population, study base, and study sample*
   - Data source
   - Sampling procedure
   - Representativeness of study base for target population

4. *Size of study and its justification*
   - Size of acceptable type II error to be considered (power!)
   - Sample size determination

- Definition of subgroups (e.g., sex, age, ethnic group, study center)
- Power calculations based on fixed sample size

5. *Selection and recruitment of study subjects*
   - Measures to avoid selection bias (survival bias, referral bias)
   - Definition of appropriate reference groups
   - Matching, if applicable
   - Access to subjects
   - Means to maximize response
   - Means to facilitate equal participation of population subgroups

6. *Outcomes, risk factors, potential confounders, and effect modifiers*
   - Operationalization and quantification of variables
   - Temporal reference
   - Choice of assessment method (e.g., questionnaire or examination)
   - Derived variables/composite scores

7. *Measurement and data collection*
   - Measures to avoid information bias (recall bias, measurement bias)
   - Type of instruments (self-completion, face-to-face, telephone)
   - Design and content of instruments (structure, layout, comprehensibility, length, sensitive issues)
   - Media (paper-pencil, computer, online)
   - Use of existing (secondary) data (e.g., hospital records)
   - Measurements and examinations
   - Coding of free text (occupation, disease, medication)
   - Biomaterials: collection, processing, shipment, and storage

8. *Data entry and data storage*
   - Database system
   - Ergonomic layout and functionality of data entry screen
   - Validation of data entry
   - Validation of coding
   - Plausibility checks (concurrent to data collection)
   - Data corrections and documentation of data cleaning (always retain raw data!)
   - Merging of data
   - Data safety (back-ups)

9. *Quality assurance*
   - Standard operation procedures (SOPs) for data collection, measurement, interview (interviewer and operation manual)
   - Training and supervision of staff
   - Monitoring of data collection
   - Assessment of response incl. minimal information on non-responders

- Audio recording of interviews
- Tasks of internal/external quality control board
- Validation studies
- Comparison with reference data external to the study
- Description of changes of original study plan
- Project diary

10. *Strategy for analysis including statistical models*
    - Strategy to analyze a priori hypotheses
    - Strategy of exploratory data analysis (if applicable)
    - Procedures for secondary analysis
    - Control of confounders
11. *Data protection and ethical consideration*
    - Informed consent
    - Anonymization/pseudonymization of individual data
    - Sensitive issues (e.g., handling of genetic data, invasive procedures, long-term storage of data or biomaterials)
    - Handling of incidental findings and feedback to study participants
12. *Timeline and responsibilities*
    - Chronological sequence
    - Project governance (e.g., steering committee in multicenter studies)
    - Scientific councils
    - Publication rules

13.3 Design

The study design governs all procedures for selecting and recruiting individuals in the study sample. It also governs strategies to avoid bias and methods for data analysis. The design depends on both the study objectives and on practical issues such as costs or availability of data. Most epidemiological studies, with the exception of clinical trials and intervention studies, are purely observational, in that the investigator cannot assign the exposure to the study subjects as in experimental settings. The definition of a study design includes a definition of both the study base and the study type. The main types of observational studies are the cohort study and the case-control study (sometimes called case-referent study). Although they have a common basis, as every case-control study has to be considered as being rooted in a cohort (see chapter ►Case-Control Studies of this handbook), the practical implications of conducting a cohort or a case-control study are quite different. A number of other designs have been used or proposed and are mostly variations of these basic types (see chapter ►Modern Epidemiological Study Designs of this handbook).

13.3.1 Study Base

Once the general objective of the study has been defined and before the study type can be decided, the investigator should identify the actual setting, or study base, which is most suitable to study the particular research question. The study base should represent not only the target population but also the morbidity experience of this population during a certain period of time (Hernberg 1992, Chap. 4). This means that both the distribution of exposures and the morbidity and mortality experience of the study base have to be comparable to the target population.

For example, if the target population consists of workers within a particular industry where a potentially hazardous chemical was used, the study base may be the 10-year follow-up of workers employed for at least 1 year in a particular plant of this industry where the chemical in question has been used during the observation period (hypothesis testing). Another example would be the general population in a country with a number of environmental and lifestyle-related exposures (target population). Here the study base could be defined as the residents of a geographical area – where these exposures are all present – in this country at a certain point in time. The association of one or more health outcomes with a wide range of exposures can then be assessed in exploratory analyses (hypothesis generating).

Ideally, a study base should allow for a scientific generalization of study results (e.g., the association between exposure and disease). This can be achieved if exposure conditions as well as potential confounders and effect modifiers can be measured within the study base. Otherwise, the particular morbidity experience of the study sample would be mostly descriptive and would apply only to the population under study.

13.3.2 Cohort Studies

The design and analysis of cohort studies is described in detail in chapter ▸Cohort Studies of this handbook. In cohort studies the study sample to be included is either the whole study base or a sample of it based on certain criteria like existing exposure information.

The complete assessment of the occurrence of the morbidity outcomes or other endpoints, for example, biological markers, during the follow-up is one of the challenges of such prospective studies. While certain endpoints can be obtained from a *passive follow-up* by record linkage with routine records of mortality or incidence using registries or administrative data (cf. chapter ▸Use of Health Registers of this handbook), others need an *active follow-up* by which the health status or the biological marker of every subject in the cohort is determined at several points in time during the observation period. In this case, the study design is also called longitudinal study or panel study (see, for instance, Wild et al. 2008a).

In principle, the assessment of the exposure is not influenced by disease status as it is recorded before the disease arises. Conceptually, all cohort studies are prospective in the sense that one measures the exposure before the disease occurrence. In practice this approach is often modified when the outcome of interest is a disease with a long induction period like cancer, so that a purely prospective observation of the morbidity would imply either a study duration of several decades or very large numbers of participants. One way to shorten this duration is to set up the cohort historically, that is, from a past point in time onward and to mimic the prospective follow-up of the cohort until present. This design is called *a historical cohort* design in contrast to the *prospective cohort* design.

When subjects are *lost to follow-up* their morbidity, outcomes and/or other time-dependent factors are not observed from that point in time onward. When such a dropout occurs, its influence can only be controlled for in the analysis if it can be assumed that the dropout is only determined by recorded factors. This means that we have to know for each missing subject why she/he is missing and in particular that the dropout does not depend on the unmeasured health outcome. In technical terms this is the "missing at random" (MAR) assumption (cf. chapter ▶Missing Data of this handbook). This assumption cannot, however, be assessed from the study data as – per definition of dropout – we cannot know his/her morbidity or other outcome. An example of this would be a study in which the outcome of interest is smoking cessation comparing different cessation strategies. If the study participants fail to show up at later interviews aimed at investigating whether the subjects are still abstinent, this may be because they have relapsed and are ashamed of this fact or because smoking is no longer a problem as they have been abstinent for a long period of time. The effect of study dropouts can then only be assessed through sensitivity studies by which one would simulate reasonable data for the missing outcome, that is, the change of behavior. In this smoking cessation example, one could, for instance, assume that all dropouts relapsed or a 90%, 70%, etc. proportion of the dropouts. Minimizing the number of study dropouts is one of the main challenges to avoid selection bias in cohort studies.

13.3.2.1 Historical Cohorts

The historical cohort design mimics a prospective follow-up of a historically defined study base. This entails that at any point in the past, it must be possible to check whether a subject satisfies the cohort-inclusion criteria as defined in the protocol. Furthermore if a subject from this cohort developed the outcome of interest, one must be sure to observe it and to determine at which point in time this event occurred. In historical cohorts, the operational definition of follow-up is the following: if, during the observation period, the event of interest occurs at a given point in time for a given subject, it must have been recorded on a routine basis, and this record must be traceable with certainty. If this cannot be assured from a certain point in time onward, the subject must be considered as lost to follow-up since then.

The historical cohort design has been the design of choice for industry-based epidemiology of chronic diseases, especially cancer (cf. chapter ▶Occupational Epidemiology of this handbook), and we shall discuss the relevant issues in this context.

Cohort Recruitment The first step in planning is to determine whether information exists by which one can be assured that the population satisfying the theoretical cohort definition can be enumerated without bias. Completeness with respect to the cohort definition is of paramount importance. For instance, if the cohort definition is "all subjects having worked on a given industrial site since 1970," a document dated from 1970 listing all those subjects must be available as well as yearly lists of all subsequently hired and dismissed subjects. Computerized files can rarely be trusted as they were usually created for administrative purposes and are often at least partially overwritten or may exclude some categories of employees. Individual files as a single source of data are insufficient. Negligence, lack of space for archives, or floods and fire may have led to the selective deletion of files. It cannot be assumed that these lacking data are independent from either exposure or health outcome. Wage lists can be considered as reliable data source as all processes related to financial transfers are usually recorded with great care. In order to reliably identify a historical cohort, at least two data sources which can be considered as independent should be available. One example would be lists from the pension scheme and salary records of the personnel department. An alternative would be a union membership list and historical documents in which yearly counts of employed subjects are given.

Case Ascertainment and Loss to Follow-Up In historical cohorts the main concern is to assure that any event of interest that occurred during the follow-up period is detected. As an active follow-up is retrospectively impossible, the tracing of subjects must rely on linkage of the cohort data with routine records or on the collection of death certificates for deceased cohort members. These routine records are usually restricted to a defined geographical area. The traditional approach of a historical cohort study is a mortality study where the cause of death is abstracted from death certificates. This approach is quite convenient in some countries that have a national death index like the USA or the Netherlands and more laborious in others where local files have to be abstracted and coded. A mortality study, however, may give misleading results if the validity for a given cause of death is low and may even not be suitable to investigate diseases with a good survival. In the latter case, linkage with disease registries or hospital records is a better approach.

Disease or mortality registries are regional in most countries (with the notable exception of the Scandinavian countries and some others like France and the USA) and sometimes linked to the place of birth or residence. When either information is incomplete or missing for some subjects or if some subjects moved out of the covered geographical area, they are lost to follow-up, that is, from that moment on an event of interest is no longer traceable with certainty. In the statistical analysis, the subject should therefore no longer contribute any person-time (cf. chapter ►Cohort Studies of this handbook).

It is important to set up procedures by which one can determine whether and when a subject is lost to follow-up, for instance, by trying to trace address and vital status of study subjects through records of residents or pension schemes. Setting up such procedures can be difficult or even impossible for specific groups. For example, historical cohorts may include foreign-born subjects who may have returned to their country of origin after enumeration. If a subgroup is identified, for which such loss to follow-up is likely but no individual tracing can be set up, the only solution might be to drop all members of this subgroup from the study at the time of last contact. Thus, in a historical occupational cohort study, one may need to consider as lost to follow-up all foreign-born employees from the date of last employment onward.

As an informed consent cannot be obtained from study subjects, recording of individual diseases or medical causes of death may be restricted by data protection laws. Specific procedures are therefore required, often involving third parties, by which the epidemiologists can link disease or death records with study data on an individual level (e.g., Wild et al. 1995a).

13.3.2.2 Prospective Cohorts

The major advantage of cohorts where individual-level data are collected prospectively is that the dynamics of the etiological process can be observed more accurately because time-dependent data such as exposures, biological markers, and confounders of interest can be measured repeatedly in temporal order with disease outcomes. On the contrary, in historical cohorts the exposure is assessed retrospectively based on historical exposure data which are not collected for the purpose of the study. Such data are limited with regard to their validity and are often not directly linked to a given individual as, for example, environmental or occupational measurements. The resulting imprecision may lead to a misclassification bias not unlike that potentially occurring in case-control studies that may mask an existing association between exposure and outcome.

Another advantage of prospectively collected data lies in the possibility to perform intermediate analyses and to include additional markers and novel instruments in subsequent follow-ups. This is typically the case in so-called multi-purpose cohorts, where multiple exposures entailing lifestyle, environmental, psychological, and sociodemographic factors are assessed together with physiological measurements and where biological materials like blood, urine, saliva, genetic material, and feces are collected and stored in biobanks for later analyses. Examples for this type of cohorts are the UK Biobank (2007; Biobank UK 2012), the Nurses' Health Study (2011), and the National Cohort (2012) in Germany.

Although there is the general belief that prospectively conducted cohorts are not suitable to study long-latency chronic diseases, that is, in general diseases with long induction periods, several very large studies like the American Cancer Prevention Studies (Garfinkel 1985), the Nurses' Health Study, or the European Prospective Investigation into Cancer and Nutrition (EPIC) study (IARC 2012) are designed to investigate endpoints like cancer and cardiovascular diseases. However, other prospective cohorts with shorter follow-up periods are primarily targeted at subclinical disorders assessed by questionnaires and functional or biological measurements. So-called birth cohorts that start with the collection of data and biological material during pregnancy and around delivery entail a particular innovation potential. They allow to study the long-term effect of intrauterine factors and possibly formative influences during infancy as well as the sequence of events leading to chronic disorders later in life (cf. chapter ▸Life Course Epidemiology of this handbook).

An inherent limitation of prospective studies is the long-term commitment needed by both the research institutions and the funding agencies. Moreover, the prospective recruitment and follow-up of populations require repeated contacts with each study subject of the cohort and are very cost-intensive. The informed consent to be given by study participants must include the permission to analyze the collected data longitudinally but also to store the contact data over an extended period of time to enable further follow-ups. It should also foresee the permission to store biological material in biobanks and to use the stored samples for analyses that cannot be defined precisely at the time of data collection because otherwise new knowledge and technological advances available only several years after their collection could not be utilized to the best scientific benefit of these materials.

The burden put on study participants may impair the willingness to participate. Thus, the participation proportion is often rather low, especially if participation entails repeated contacts or an extensive examination protocol. For instance, participation varied between 22% and 38% in the German centers of the EPIC study (Boeing et al. 1999). Limited information (Goldberg et al. 2001) tends to show that participants are generally in better health than non-participants. Thus prospective cohorts are not fully representative of their target populations. Fortunately, representativeness is not a key issue in such studies as long as the loss to follow-up is limited. In prospective cohorts, loss to follow-up occurs either due to actual loss of contact or more often due to the refusal of continued participation. An unpublished survey by Moulin (personal communication) of all large ongoing prospective cohorts in France suggested that loss to follow-up can only be reduced by regular contacts, regular information about the study results to the participants, and, if possible, presence in the media. Researchers should incorporate elements that involve a benefit to study participants themselves, for example, by providing useful feedback about a subject's health status or by providing individualized advice for health promotion. When planning prospective studies, enough resources should therefore be allocated not only to the recruitment of the subjects (i.e., mailing of questionnaires and reminders) but also to the communication, both with the media and with the study participants. Also, new media like the Internet and mobile devices may become important avenues through which participants can be motivated and cohorts can be maintained (see chapter ▸Internet-Based Epidemiology of the handbook).

13.3.3 Case-Control Studies

In a case-control study, subjects who have contracted a given disease are gathered from the study base and are contrasted to a sample of control subjects drawn from the same base. Exposure histories are then collected from cases and controls. The basic principle of a case-control design (also called a case-referent approach as in Miettinen (1985)) is closely related to that of a cohort study (see also chapter ▸Case-Control Studies of this handbook) as it can be considered as an instantaneous condensation of a cohort. It resembles the morbidity experience of a cohort – represented by the cases – without the need to observe the full person-time at risk. Instead of comparing the disease incidence between exposure groups, it compares the exposure between controls representing the population at risk and cases, that is, the exposure history that was accumulated by both groups over the person-time at risk is assessed retrospectively.

According to the most frequent definitions, a case-control study conducted within a dynamic population may be *population-based* or *hospital-based* depending on the selection procedure of cases and controls. When the study base is a previously enumerated cohort, the case-control study is often called *nested within the cohort*. Each of these types has different practical implications which we will detail below. In most cases it is recommended to include only patients who were newly diagnosed, in particular when the case fatality rate is high, in order to avoid selection effects.

13.3.3.1 Population-Based Case-Control Study

In a population-based case-control study, the cases are all patients newly diagnosed with the disease during the study period in a defined population, for example, those who live within a country or a geographical region. The controls are sampled from this same population. In this design, the study base, that is, the population living in this country or region during the study period, is precisely defined a priori before the start of the study.

Case Recruitment

A researcher may choose a population-based case-control design if all new cases of the disease (or at least a representative sample of all cases) can be identified in the study base. As it is essential to ensure completeness of case finding, a disease registry may be helpful for identifying the cases. The existence of a disease registry may also guide the decision for a given study region. However, the value of a registry may be limited if there is a substantial time lag between diagnosis and registration, particularly if the disease is rapidly fatal and collection of data from the respondent is necessary. Thus, an efficient procedure for case identification and recruitment has to be set up by the research team. For example, a case-finding network may be organized including all hospitals, clinics, and pathology departments in the study region to identify and interview the cases. This should also be extended to nearby areas if some of the diseased persons in the source population defined by residence in a given geographical region may be diagnosed and hospitalized outside that region. It is also strongly recommended to perform an active search of the patients, by organizing periodic retrieval of hospital records, rather than to rely on a passive notification of the cases by the medical staff of the clinics or the hospitals.

Selection of Controls The controls should be selected at random from the same source population giving rise to the cases. Ideally, the probability of selecting a particular control subject should be proportional to the amount of time that he/she contributes to the study period or to person-time at risk (Rothman et al. 2008, p. 116). For example, if a subject moved out of the source population at half the study period, he/she should have only half the probability of being selected as a control than a subject who stayed in the source population during the entire study period. To be eligible, a control subject should belong to the study base at the date of diagnosis of the index case. Controls who recently moved into the source population and are chosen to match cases diagnosed several years earlier should be excluded since they do not belong to the study base. Excluding controls who have recently moved into the population reduces the problem, but does not solve it, since people who have moved out of the base will still be missed.

An incidence density sampling of the study base can avoid these problems. Density sampling can be achieved, for example, by selecting the controls at a steady rate throughout the study period proportional to the number of cases. In practice, the protocol may define several points in time, for example, once a month or once a year, where controls will be selected from the population present in the study base at that time. In a design where the cases and the controls are matched individually, it is also possible to use sampling sets of possible controls, one per case, composed of all persons present in the source population and at risk of contracting the disease at the time of the case's diagnosis. The desired number of matched controls is then drawn at random from each of these risk sets.

Selection of Population Controls Based on a Listing of Individuals

To be feasible, these procedures of selection of controls necessitate not only a fully enumerated study base but also regular updates of this population, to take emigration and immigration during the study period into account. In Scandinavian countries, study investigators may rely on central population registers to select controls using a simple random sampling at regular intervals during the study period (see also chapter ▸Use of Health Registers of this handbook). More often, however, a complete population register, including the identification of individual members by name and address as well as stratification variables such as sex and date of birth, will not be available, and other methods of selection of controls must be chosen.

In the absence of a population register, the researcher may use other *lists of individuals*, such as lists of municipality residents, electoral lists, and listings of health insurance members. As long as such lists cover 100% of all residents in the study region, they are as useful as population registers. In other cases, using these lists for selection of controls may, however, introduce bias if the probability for an individual in the study base to be listed is related to the exposure of interest. Telephone books would not be suitable for a study on cancer in relation to an occupational chemical, for example, if phone numbers of highly educated and less exposed individuals are less frequently published in the directory than phone numbers of subjects of other socioeconomic categories. Persons not registered in electoral rolls may also differ from those listed and may not include immigrant workers. Municipal lists may not be updated regularly. Besides completeness of the list, the possibility of tracing individuals based on the information provided (name, address, phone number, etc.) must also be considered. One may cross-check whether the cases are also listed on the roster from which the controls are drawn. The analysis should then exclude the cases not listed, such as those who are not citizens when using electoral rolls.

Selection of Population Controls Without a Listing of Individuals

Other sampling schemes may be used for selecting controls when no list of residents is available. Multistage random samplings starting with sampling of dwellings or based on random digit dialing procedures are commonly used. The controls then have to be selected within each household.

Neighborhood Controls This method implies a two-stage sampling, with a random sampling of households followed by the selection of an eligible individual within the selected residence. Households sampling may be conducted from a roster of residences, obtained, for example, from census data. When a roster of households is not available, controls may still be selected from residences in the case's neighborhood. Starting from the case's residence, the interviewer may follow a predefined procedure for selecting a household, by means of a map, aerial pictures, or by a systematic walk algorithm starting at the index household (Wacholder et al. 1992b). This sampling method implies that the controls are individually matched to cases on place of residence. To avoid bias, the interviewer should not be given the flexibility to choose which house to select, rather a simple and unambiguous algorithm for selecting households should be developed to remove the possibility of interviewers avoiding certain areas. A potential problem of neighborhood controls is overmatching on the exposure due to similarities between cases and controls living in the same neighborhood.

Random Digit Dialing Random digit dialing can be used to select population controls when no population roster exists and when almost every household has a landline telephone. Random digit dialing generates telephone numbers, and it does not rely on a telephone book where new or unpublished phone numbers are not listed. Several variants of the standard method exist (Waksberg 1978). Briefly, a phone number is created using the first numbers, including area code, of working telephone numbers provided by the telephone company, which are then completed with random numbers. The number is dialed a predetermined number of times. The first contact with a household member is used for screening and to obtain a census of the household. Based on the responses and a predetermined sampling scheme, eligible subjects are selected to be controls. These individuals can be interviewed by telephone directly, or they can be contacted afterward for an in-person or a telephone interview.

Random digit dialing is not appropriate if the coverage of landline telephones is low, which is an increasing phenomenon since mobile phones conquered the market. Other problems associated with random digit dialing include residences that can be reached by more than one phone number or if more than one person in the household is eligible to be a control, since it may lead to different selection probabilities of the controls. It should also be realized that random digit dialing is an expensive and time-consuming procedure, particularly when targeting subgroups of the population, since a large number of phone calls may be necessary before the desired number of eligible controls are recruited. Non-response and refusal is an additional problem, and it may not be possible to have an exact estimate of the participation proportion. It is recommended that the distribution of the final sample is compared according to key variables such as age, sex, or socioeconomic status to an expected distribution obtained, for example, from the last census.

The random digit dialing has been used successfully in a large number of studies, but the widespread and sometimes exclusive use of mobile phones or Internet-based phone numbers causes further problems as these numbers are not defined regionally in many countries.

13.3.3.2 Hospital-Based Case-Control Study

In hospital-based case-control studies, the cases are patients admitted to a given hospital or clinic who were diagnosed and hospitalized with the disease during the study period. The controls have to be selected from the population from which the cases arose, that is, the group of individuals who would be treated in this hospital if they had developed the disease. Because the study base is not easily identified as the catchment areas of hospitals are not precisely delineated, a random sample of controls can hardly be selected directly from this population. Instead, it is usually more practical to select controls among patients admitted for other diseases to the same hospital, representing a non-random subset of the study base. An appropriate group of control patients should have the same referral pattern to that hospital as the cases, so that the controls have the same probability as the cases to be admitted to this hospital if they had the disease under study.

The possibility of selecting hospital controls rests on the assumption that they are representative of the exposure distribution in the study base. This assumption is reasonable if the hospital admission of controls is not associated with exposure. In a study on smoking and lung cancer, for example, patients admitted for smoking-related diseases like bladder cancer or myocardial infarction have a higher probability of being smokers than it is the case in the study base. Their inclusion as controls would therefore lead to an underestimation of the strength of the association between smoking and lung cancer. The appropriate choice of hospital controls becomes difficult, if the knowledge about the diseases that are caused by the exposure(s) of interest is limited. Two different strategies have been proposed to deal with this problem. According to the first strategy, a single disease that – according to current knowledge – is unrelated to the exposure(s) of interest is selected. In cancer studies often colon cancer has been selected for controls because it is not or only weakly associated with many agent exposures that have been studied (Lynge et al. 2005). The advantage of this approach lies in the fact that cases and controls suffer from a disease of comparable severity which may reduce a possible recall bias. Also, because the disease eligible for controls is well-defined, subsequent evidence of a specific – previously unknown – association with an exposure of interest may be taken into account by a correction factor or for the quantitative assessment of the selection bias in a re-analysis of the case-control study.

The second strategy tries to minimize the risk of biased risk estimates due to a – yet unknown – relationship between the exposure(s) of interest and the disease eligible for controls by dilution. Here all diseases that are unrelated to the exposure(s) are eligible as controls, but none of them should dominate the control group. For example, in the European case-control study of head and neck cancers, none of the diseases in the control group were allowed to exceed a proportion of 30% (Lagiou et al. 2009).

13.3.3.3 Case-Control Study Nested Within a Cohort

When the study base consists of an enumerated cohort with recorded entry and exit times, a case-control study may be initiated by drawing cases and controls from this cohort as the source population. This is called a "case-control study nested within a cohort" which requires an existing cohort from which the controls can be selected. In a matched design, for example, a set of possible controls can be constituted with all non-diseased individuals in the cohort at the time of each case's diagnosis, the so-called risk set. One or several controls may then be drawn at random from each set (see chapter ▸Modern Epidemiological Study Designs of this handbook).

An advantage of this design consists in the efficient and cost-saving use of biological samples that were collected within the framework of the cohort study because the expensive laboratory analyses can be restricted to those who contracted the disease and a sample of cohort members serving as controls. Another reason for embedding a case-control study into an existing cohort is the possibility to enrich the data, for example, by performing an in-depth exposure assessment or by retrospectively collecting confounder information that is unavailable for the whole cohort. This was done in the case of a historical cohort study of workers exposed to bitumen fumes which showed inconsistent results with regard to lung cancer mortality. As some positive associations might have been due to confounding by smoking while some inconsistent exposure-effect relationships might have been due to misclassification of exposure histories, a nested case-control study was performed including an individual exposure assessment and the collection of smoking histories for each subject (Olsson et al. 2010).

13.3.4 The Study Base Principle: Selection, Exclusion, and Resulting Bias

The study base principle, the first of three principles Wacholder et al. (1992a) developed for the selection of controls in case-control studies that also applies to the design of cohort studies, simply states that "cases and controls should be representative of the same base experience" (see also chapter ▸Case-Control Studies of this handbook). *Representativeness* for the target population is not a conditio sine qua non in studies of the association between an exposure and a disease. Thus, for the aim of scientific inference of the relation between an exposure and disease, any exclusion or inclusion criteria are valid as long as they apply equally to cases and controls. However, if an exposure-disease association is observed in a highly selected group, the external validity of this association is deemed to be questionable.

Wacholder et al. (1992a) identified the following reasons for which exclusion criteria can be applied:

- Inconvenience: Subjects of a given subgroup might be hard to reach. Failure to exclude a priori such a group may lead to a very poor response proportion and an a posteriori exclusion, with the corresponding waste of resources.

- Anticipated low or inaccurate responses, for example, of subjects who do not speak the language of the interview sufficiently. Failure to exclude a priori such a group may yield non-interpretable data.
- Lack of variability of the exposure: If one intends to set up a cohort investigating the dose-response effect of a potential occupational carcinogen like cobalt salts, inclusion of a large number of workers from industries who do not use these chemicals makes little sense although including a small group of unexposed may still be justified for stabilizing the baseline category.
- Subjects at increased risk of disease due to other causes: In a prospective study investigating environmental effects on asthma, subjects at high asthma risk due to their occupational exposure (e.g., bakers) should be excluded because cases are likely to be attributable to the occupational exposure and therefore may not contribute to the understanding of other risk factors.
- Combination of the above: In a historical occupational cohort study based on a factory, short-term employees may be difficult to track, their exposure is likely to be determined much more by previous or subsequent work, their cumulative exposure within the company is bound to be low, and they may be at increased risk for many diseases as they constitute a group of socially unstable workers who are likely to have other risk factors. It is thus standard in such settings to exclude short-term workers. This has, however, the consequence that any given subject of the study base contributes person-time only from the date on when she/he has reached the minimum employment duration.

As said above, the comparison of a highly selected study sample with the target population may be misleading. The so-called *Healthy Worker Effect*, by which is meant that a series of extraneous factors lower the observed mortality among employed workers in an occupational cohort, is an example of such a selection bias. A simple comparison of the mortality of a cohort with population mortality rates as expressed, for instance, through the standardized mortality ratio is thus of limited validity. Of course, as discussed above, internal comparisons of exposure groups are still valid but may lack the necessary power for useful scientific inference except in very large cohorts.

Another consequence of very restrictive eligibility criteria in a cohort or case-control study is the reduced variability of exposures. This may reduce the ability of a study to detect an existing exposure-effect relationship if the spread between low and high exposure is small. It may also limit the ability to detect any effect modification: If the effect of smoking were to amplify the effect of an environmental exposure, restricting the study base to non-smokers may lead to a spurious absence of effects.

In some instances, representativeness is an important requirement. If the study aims to assess attributable or absolute risks in the target population, the study must be either exhaustive or representative. Otherwise, external information must be retrieved with respect to the fractions of the strata in the study sample relative to the target population.

13.3.5 Choosing Between Epidemiological Designs

Many other etiological designs like case-cohort designs, case-only designs, two-phase sampling, counter-matched designs, and multilevel designs to cite just a few have been proposed in the epidemiological literature (Wacholder et al. 1992c; chapter ▶Modern Epidemiological Study Designs of this handbook). These designs usually combine elements of the case-control and the cohort design either by stating additional hypotheses (case-only design) or by making use of additional data like in two-phase designs (see, e.g., Breslow and Chatterjee 1999; Pohlabeln et al. 2002).

Another design is the so-called ecological design, in which the units of investigation are groups of people rather than individual subjects. The groups may be classes in a school, factories, cities, or administrative areas within a country. The only requirement is that a measure of the exposure and disease distributions is available for each group. Because the data in ecological studies are measurements averaged over individuals, the degree of association does not reflect the association between exposure and disease among individuals (so-called ecological bias, see Greenland and Robins (1994)). Thus, while ecological studies can be useful for detecting associations of exposure distributions with disease occurrence, that is, for hypothesis generation, such a design should not be used for etiological investigations and quantification of exposure-disease relationships.

Another possible design consists in selecting a cross-section of the study base with no time dimension (cross-sectional study). Here both exposure and disease status are collected simultaneously at one point in time. It is therefore not the incidence but rather the prevalence of the disease that is investigated, so that it is usually impossible to assess whether the exposure actually preceded its presumed effect on health. Moreover, cross-sectional studies are particularly prone to selection bias as diseased subjects may have left the study base or may be less willing to participate. A longitudinal observation of exposure and disease in a study that is the paradigm of both the cohort and the case-control design is better suited for answering etiological questions than a prevalence study. Possible exceptions are diseases with short induction periods or exposures that cannot change such as genetic characteristics or other invariable personal characteristics (see Rothman and Greenland 1998, p. 75). If, however, a prospective cohort design is not feasible for financial reasons or simply because it is impossible to follow up a large enough group of subjects, the cross-sectional design can, despite its above-mentioned limitations, provide important information especially if targeted on non-fatal health outcomes (see Wild et al. 1995b) or if a number of different exposures and health outcomes are to be investigated simultaneously. Moreover, a cross-sectional study may serve as the starting point for a subsequent follow-up of (non-diseased) study participants which will then evolve into a multipurpose cohort. An example for this is the US National Health and Nutrition Examination Survey (NHANES) which formed the basis for the NHANES I Epidemiologic Follow-up Study (NHEFS) (Centers for Disease Control and Prevention 2010).

Table 13.1 Strengths and limitations of different observational study designs

	Study design		
Investigation of . . .	Cross-sectional	Case-control	Cohort
Rare diseases	–	+++++	–
Rare causes	–	–	+++++
Multiple endpoints	++	–	+++++
Multiple exposures including confounders	+++	++++	(+++)[a]
Temporal sequence of exposure and disease	–	(+)[b]	+++++
Direct measurement of incidence	–	(+)[c]	+++++
Long induction periods	+	++++	(+++)[d]

Suitability of study design: +++++ highly suitable, ++++ very suitable, +++ suitable, ++ moderately suitable, + limited suitability, – not suitable

[a] If prospective (multipurpose cohorts)

[b] If nested in a cohort

[c] If population-based, combined with an incidence study

[d] If historical

The main study types to assess etiological relationships remain the cohort and the case-control designs. Choosing one or the other design depends on a number of issues among which the incidence rate of the disease of interest and the prevalence of the exposure of interest are prominent. Table 13.1 summarizes strengths and weaknesses of the major study types.

On the one hand, if the disease is rare (for instance, a rare cancer as testicular or brain cancer), a cohort approach would necessitate huge numbers of participants and a long follow-up to identify enough cases to make useful inferences. A case-control study is the more efficient approach for rare diseases. Another situation where the case-control design is the preferred approach is if the aim is to generate hypotheses concerning various exposures in relation to a given disease. Determining possible occupational origins of laryngeal cancer is such an example.

On the other hand, if the exposure is rare and restricted to specific subpopulations or if one is interested in all possible disease outcomes of a given exposure, a cohort study, either historical or prospective, is the most efficient choice. Examples of the former are the studies of the carcinogenic effects of hard-metal dusts among manufacturers of hard-metal tools (Moulin et al. 1998a) or of bitumen fumes in asphalt and road-paving workers (Boffetta et al. 2003). An example of the latter is the study of the health consequences of hormonal replacement therapy in postmenopausal women (Nurses' Health Study 2011).

Another aspect that can influence the choice of a design is the induction period of the disease (although it may depend on the exposure). For long induction periods, a prospective cohort study is clearly not the design of choice as this would imply waiting a long time for a sufficient number of events to occur. This problem can be circumvented by the historical cohort design. Most historical cohorts are focused on cancer, a long-induction disease per se. The choice between a case-control and a historical cohort study is then dependent on whether (or not) a historical cohort can provide an answer to the research question.

Other issues which might influence the choice of a given design include the precise scientific aim of the study. Direct estimation of the incidence of disease would require enrolment of the target population that ideally needs a cohort design or at least a population-based case-control approach. If one is, however, interested in the precise temporal sequence, as, for instance, in the study of the evolution of CD4+ cell numbers in HIV-infected patients (Kaslow et al. 1987), cohort studies are virtually the only available design which may, however, be supplemented by a nested case-control study.

The final choice, once a theoretically optimal design has been determined, depends on the actual feasibility of a given study type as well as on the practical terms of access to the data as well as the time and costs involved.

13.3.6 Statistical Power and Sample Size

As mentioned in the first section, any study protocol should include a calculation of the sample size to detect a predefined effect of the exposure for a given statistical power and a predefined probability of the type I error (significance level). Alternatively, if the available sample size is limited, a power calculation should inform about the effect size that can be detected. When computing the statistical power or sample size, the need for building subgroups, based on either exposure levels or confounder strata, should be taken into account.

If several exposures which differ in their prevalence and/or in their expected effect size are to be investigated, then a table displaying the necessary sample size for different combinations of effect sizes and exposure prevalences can be very helpful to find a reasonable compromise between power and sample size. An example is given in Table 13.2 for a case-control study. For example, if you want to detect an odds ratio of 2.0 at a significance level of 5% with a power of 80%, you would need to include at least 449 case-control pairs in a 1:1 matched study, if the prevalence of the exposure in question is 5% in the study base. A similar approach can be used for a cohort study where a given exposure is to be investigated with regard to outcomes with different incidence rates.

If the study addresses more than one hypothesis, the power calculation should be done for the main hypothesis. If, however, several hypotheses have to be evaluated simultaneously based on a statistical test, that is, in case of a so-called multiple

Table 13.2 Number of case-control pairs required to provide an 80% power to detect a given relative risk at the 5% significance level

Odds ratio	Proportion of population exposed			
	1%	5%	10%	20%
1.5	6,672	1,415	763	448
2.0	2,087	449	246	150
2.5	1,114	243	135	84
3.0	732	161	91	58

testing problem, appropriate methods have to be applied to account for multiplicity. The simplest but also most conservative approach is the Bonferroni correction where each statistical test out of, say, k tests is not performed at significance level α, but at a significance level of size α/k to control the family-wise error rate. In case that thousands of hypotheses have to be tested as, for example, in genome-wide association studies, this approach is of course no longer feasible, and more sophisticated methods have to be applied (see chapter ▸Statistical Methods in Genetic Epidemiology of this handbook).

In practice the choice of the sample size (or to be exact the size of the study sample) is a compromise between what one would ideally be able to detect and practical limitations. A common limitation can be an insufficient number of study subjects, either because the incidence of a disease is low as in the case of rare cancers or because the exposure is rare. An example of this would be if we were to study the interaction between a rare genetic variant (prevalence <0.01) and a rare environmental exposure. If the latter has a prevalence of 5%, less than 5/10,000 of controls would show both features so that the minimum sample size to investigate such an interaction would be in the tens of thousands.

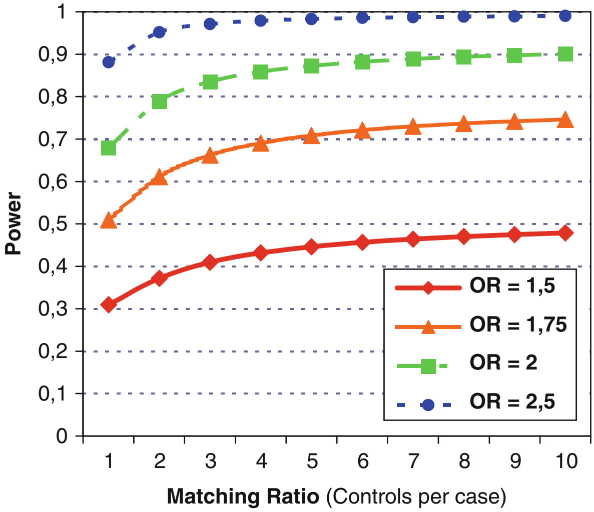
In general, a statistical power of 80% is considered a reasonable power to achieve objectives of an observational study investigating etiological associations, and any power below this arbitrary figure may be considered too low. In case that one needs to reduce the risk of failing to detect a certain effect that is truly present – as it might be the case in clinical trials – one needs to increase the statistical power which of course inflates the necessary sample size. If methods of prior assessment, either formal or intuitive, suggest that the study will be too small to be informative, there are several options:

One can reduce the level of ambition. For instance, instead of choosing a statistical power to detect an odds ratio of 1.3, the detection limit can be raised to 1.5. The drawback of this strategy is of course that if no effect of the exposure is observed, risks below the level for which the study is powered cannot be excluded. Thus, if the main interest lies in exposures with presumably low effect sizes, then such a strategy is not feasible.

If the number of cases is the limiting factor in a case-control study, the study power can be increased by selecting several controls per case. Up to four controls per case lead to a reasonable gain of statistical power, while further increases of the case to control ratio have only a marginal effect (Fig. 13.2). One may also extend the recruitment period to enhance the number of incident cases. This usually entails that the results will become available later. An analogous strategy would be the extension of the observation period (follow-up period) in a cohort study to increase the person-time at risk and thus the number of events.

Another strategy consists in the combination of several smaller studies in one multicenter study. When organizing multicenter studies, one should be reasonably sure that the gain in sample size is not offset by between-center differences in exposure circumstances and assessment or differences in case ascertainment. Another issue in this case is that the harmonization of centers has its costs too. An advantage, however, may be the broader range of exposure levels which may lead

Fig. 13.2 Statistical power of a case-control study for various matching ratios of m controls per case to detect different odds ratios (OR) for a number of 100 cases with $\alpha = 0.05$



to a better chance to detect an exposure-effect relationship. If there is no feasible compromise, one has to abandon the project. It can be considered unethical to undertake a study which would not add to the general knowledge but costs money and resources which could better benefit other research. Finally one can increase the statistical power by making use of external information through the aforementioned innovative design (see, e.g., Wild et al. 2008b).

Statistical power is, however, not the only criterion by which to judge the appropriateness of a study. In addition to the power of a study, the relevance of the research question and the expected knowledge gain have to be justified in light of the presumed qualitative and quantitative dimensions of the effect. Also the burden and implications for study participants have to be balanced against ethical concerns and the expected scientific merit.

13.4 Measures of Disease Outcome and Exposure

13.4.1 Measurement and Classification of Exposure

13.4.1.1 Choice of the Exposure Measure

Epidemiological studies are designed to assess the impact of an exposure or of a preventive measure on the development of a disease. The sources of error and the ways in which exposure and disease are assessed are quite different, and thus the mechanisms by which errors arise are different as well.

The range of exposures of interest in epidemiology is broad (Savitz 2003). Exposures include exogenous agents such as drugs, diet, and chemical or physical hazards present in the environment; genetic attributes that affect ability to metabolize specific compounds; stable characteristics such as height or eye color;

physiological attributes such as blood pressure; life habits such as physical exercise or tobacco smoking; mental states such as depression; and social environment. This wide range of exposures corresponds to many different methods for measuring exposure (White et al. 2008; chapter ▶Exposure Assessment of this handbook).

The measurement method depends on the exposure under study, the required precision and accuracy, the availability of existing records, the ability of a subject to recall an exposure, cost of a given method, etc. The study protocol should describe the operational approach chosen for exposure ascertainment and its rationale. The accuracy of an operational approach is best described in relation to the method of ideal measurement, that is, the gold standard (if it exists) which is usually not feasible in epidemiological field studies.

The ultimate goal of exposure assessment is to quantify the exposure that directly contributes to the etiological process under investigation. Ideally, an exposure assessment should focus on the biologically effective exposure. Most often, however, this goal cannot be reached by an operational exposure indicator. One reason is that the biological mechanism by which exposure might cause disease is often unknown. For example, in studies on the potential cancer risks associated with exposure to extremely low-frequency (ELF) magnetic fields, it was not clear whether the most relevant exposure indicator with respect to etiology should be an average exposure over the entire life or over the most recent period before cancer diagnosis or if it should be a measure of peak exposures over a certain threshold or during particular time windows. This problem could be partially overcome if a temporal exposure profile can be assessed, to calculate different exposure indices, where each of them is then analyzed with regard to its association with a disease. Another reason is that it is often impossible to obtain the data that reflect the biologically effective exposure. For example, the persistent organochlorine pesticide dichlorodiphenyltrichloroethane (DDT) and its metabolite dichlorodiphenyldichloroethylene (DDE) were suspected to be causally related to breast cancer risk (Wolff et al. 1993). If we assume that the biologically effective exposure is the level of DDT/DDE present in breast tissue in the time window 5 to 15 years before cancer diagnosis, it is obvious that this exposure cannot be measured directly in a case-control study.

Instead, different exposure metrics could be used to study the relationship between DDT/DDE exposure and breast cancer (Savitz 2003), including environmental levels measured in the area of residence at the time of diagnosis (taking into account the residential history of the subject), present-day blood levels, or blood levels measured in the etiologically relevant period using serum specimens drawn in the past and kept in a biobank. Clearly, these exposure measures are not equivalent as they correlate differently with the biologically effective exposure. Using environmental exposures as a surrogate exposure indicator is probably ineffective, since the observed breast cancer risk may be biased toward the null due to non-differential misclassification and thus fall below what can be detected. However, if blood levels of metabolites are strongly correlated with breast tissue concentrations even after several years, an association with breast cancer may still be observable. Therefore, although blood levels are not the ideal measure, certain metabolites measured in blood may serve as an indicator of past exposure if the half-life of the agent is sufficiently long (Flesch-Janys et al. 1998). The choice of an informative indicator of exposure requires knowledge about chemical and biological processes and adequate models describing the relationship between the indicator and the exposure of interest.

All laboratory data are subject to error due to imprecise measurement. However, the conceptual error by which the measure obtained relates to the exposure of interest is often of much greater importance. In a study on the effects of microbiological contamination, measurement of viable colonies may be only of marginal interest if these bacteria are not the pathogenic ones. Bacterial endotoxin has been shown to be the more relevant exposure with respect to lung diseases (Rylander 2002).

Another challenge in exposure assessment is that often different types of exposure coexist and their individual effect cannot be separated from the effect of accompanying exposures. The level of exposure categorization must reflect scientific hypotheses. For instance, a nutritional epidemiology study investigating the role of coffee in miscarriages could consider two relevant categorizations. The role of caffeine itself would be investigated by grouping all of its sources including tea and caffeine-containing medications, whereas the role of constituents of the coffee other than caffeine would be investigated by grouping caffeinated and decaffeinated coffee.

13.4.1.2 Temporal Aspects

Some exposures are constant over time, such as genetic constitution, but all exogenous exposures such as diet and chemical pollutants vary substantially over time. In addition to the identification of a biologically relevant exposure, it is necessary to identify an etiologically relevant time window during which the exposure may be related to disease occurrence so that the data collection concentrates on etiologically relevant time windows.

It has been long recognized that many diseases such as cancer appear a long time after they have been induced. The time between the beginning of the exposure and the clinical manifestation (detection) of a disease is often called the *latency* period, although in a strict sense, latency period refers to the period between disease initiation and detection. It serves as a surrogate for the biological *induction* period in epidemiological studies, that is, the time between the beginning of the exposure and the initiation of the disease process. Often, the onset of exposure does not necessarily result in immediate induction, as for example, in the case of cancer. Epidemiological studies should allow for the fact that diseases with long induction periods that were contracted immediately after the exposure may not be attributable to this exposure. In the statistical analysis of such data, the usual practice is to shift the exposure by a given *time lag* which is typically about half the usual induction period, that is, exposure during recent years is ignored.

The consequences for planning are that whenever the presumed induction period is long, the investigator must include a sufficient number of subjects for whom the exposure reaches back far enough. Whenever the expected effects of exposure are of a short-term nature as, for instance, a reversible genotoxic effect as assessed by a comet assay, it is important to precisely assess the relevant short-term exposure.

The temporal pattern of the exposure is of interest in itself. If peak exposure is the metric that is most strongly associated with disease risk, then its effects are likely to be much more important when the exposure is highly variable over time than in circumstances in which the exposure is virtually constant. For such presupposed effects, the exposure metric of primary interest may be the number of peak exposures. If we, however, assume that the exposure acts through a cumulative damage, the total cumulative exposure is the adequate metric to express its effects. If the effect of an exposure on a disease is reversible, the cumulative exposure may be insufficient, and time since cessation of exposure needs to be considered in addition.

13.4.1.3 Sources of Exposure Data

Various sources of exposure data and their characteristics are described in White et al. (2008). The following section summarizes some key issues covered in this book (see also chapter ▸Exposure Assessment of this handbook).

Questionnaires A prominent and often the only way to assess an individual exposure is by questionnaire. The main types of instruments are self-administered questionnaires or personal interviews, administered by either telephone or face-to-face. Traditionally, self-administered questionnaire was a synonym for a simply structured paper-pencil instrument. These are more and more replaced by web- or computer-based questionnaires.

Interviews and computer-/web-based questionnaires allow avoiding errors and missing values. In particular, complex skip patterns can be used where specific questions are only presented after some trigger information has been provided by the respondent. This has been used in occupational exposure questionnaires in which job histories are obtained and specific questionnaires are only used for a limited number of jobs and/or tasks (Ahrens et al. 1993; Ahrens 1999). However, the interviewer may increase error if he/she influences the subject's responses by his/her appearance, manner, method of administration, etc. Intensive training of field staff, strict adherence to the wording of each question, and provision of fixed probing questions for the interviewer support standardization of the interview. By training interviewers in neutral interviewing techniques and by standardization of questionnaire wording and administration, the likelihood that the interviewer's personal attitudes will affect the responses is much reduced.

The use of paper and pencil for personal interviews has mostly been replaced by computer-assisted telephone interviews (CATI) or computer-assisted personal interviews (CAPI). By this technology, replies are entered electronically on the spot, and in-built plausibility checks provide instant control of possible errors. However, since validation of data entry is not possible, typing errors are not fully controllable. As the use of computers and the access to the Internet have become more and more common, their use for self-administered questionnaires (and tests) will increase in the future. They will also allow an easy and intuitive use by respondents, particularly if clearly arranged and combined with touch screen monitors. They may also help to increase the willingness of study subjects to participate if they are offered to complete the questionnaire at home or at work via the Internet. However, the validity and reliability of such instruments need to be assessed (see chapter ▸Internet-Based Epidemiology).

The main advantage of a self-administered questionnaire is its relatively low cost. Armstrong et al. (1994, p. 44) concluded that “there appears to be little difference between these methods (subjective recall of exposure collected through face-to-face, telephone or self-administered questionnaires) with respect to the validity of the data obtained. (. . .) Face-to-face interviews are the dominant approach and are clearly best for the collection of large amounts of complex data. However, where subjects are widely dispersed and the questionnaire can be kept comparatively brief, telephone interviews can be favored. Self-administered questionnaires should be considered for low budget studies for which small amounts of reasonably simple data are required.”

In setting up a questionnaire, many sometimes contradictory issues arise. While obviously more details can theoretically be assessed if it is longer, long questionnaires take more time and (especially among diseased subjects) may become burdensome and lead to fatigue effects. A questionnaire has to be well arranged and should be comprehensible for less educated people. Questions like “Have you been exposed to bischloromethylether?” should be avoided. Sensitive issues (religion, sexual habits, alcohol consumption, etc.) should be avoided if they are not central to the study and should be given careful consideration if necessary in order to avoid withdrawal of the respondent. Embarrassment of subjects and biased replies can be reduced by using self-completion instruments, by explaining why such a question is asked, and by placing the question toward the end of the questionnaire (see also chapter ▸Epidemiological Field Work in Population-Based Studies of this handbook).

A questionnaire should always be tested before use. The use of validated instruments is of course desirable. A large number of publications exist on validating existing questionnaires (see, for instance, Rouch et al. (2003) or Bogers et al. (2004)). It is probably a good strategy to choose, whenever possible, an already existing questionnaire that has been validated in settings and target groups comparable to its intended use. If one decides, nevertheless, to adapt a questionnaire for a given study, a pretest should be carried out to investigate its properties, notably feasibility, clarity, and reproducibility, so that necessary adaptations can be made before applying it to a broader group. If a completely new instrument has to be designed, a validation study should be taken into consideration to investigate its properties as, for example, validity and responsiveness to change, that is, its ability to reflect changes in behavior or subjective symptoms (Bogers et al. 2004).

Diaries Diaries refer to detailed prospective records of exposure by the subject. As such they can be used neither in case-control studies nor in historical cohort studies. This method has been used in many contexts among which their use in nutritional epidemiology for measuring dietary intake is prominent (cf. Andersen et al. 2004; chapter ▸Nutritional Epidemiology of this handbook). Ongoing monitoring of symptoms of a disease is another application of diaries (cf. Goebel et al. 2002).

Armstrong et al. (1994, p. 219) concluded, “The use of diaries may be highly accurate method of measuring present common behaviors. The limitations of diaries, in comparison with interview methods, are the greater burden on subjects, which may lead to poorer response rate and the greater cost for subject training and for coding the data. The accuracy of diary information can be enhanced by use of multiple diary days spread over a sufficient time period, and by careful training of subjects and coders.”

Records of Exposures Historical records may be a valuable and sometimes the single source of early exposure data. Usually such data are recorded for administrative purposes or to account financial claims as in the case of health insurance data. Two types of records can be useful for exposure assessment. A first type of records contains information on the individual study subjects, for instance, prescriptions, diagnoses, and treatments contained in medical records but also social or occupational data contained in population registries. A second type of records provides exposure data only for groups of people. Examples are environmental exposures assessed by ambient air monitoring or noise measurements or descriptions of histories of industrial processes in occupational epidemiology. The main drawback of such group level data is the inherent inaccuracy in ascribing exposures to individual subjects and the difficulty to assess the validity of the historical data. The primary advantage of records is that they provide prospectively recorded information that is not influenced by health outcomes. Exposure assessment based on historical records is robust against information or recall bias. For example, use of pharmacy records in a case-control study of prescription drug use could replace a questionable recall.

Biological Measurements and Biobanking The collection of blood, urine, saliva, feces, or tumor tissue is common practice in epidemiological studies to assess medical endpoints, biological effects as early endpoints or predictors of later disease, genetic polymorphisms, or markers of exposure. In principle, measurements made directly in the human body might be considered the ideal approach to assess exposure for etiological studies because they do not depend on a subject’s ability to recall a previous exposure to a specific agent and because they are not biased by subjective judgment of a respondent.

In practice, however, a number of problems exist. A first problem is to identify the correct metabolite and the appropriate point in time that is relevant for the presumed biologically effective dose. At the same time it should always be strived for the least invasive method to obtain the marker of interest. Particularly in population-based studies, more invasive methods may reduce the willingness to participate, that is, if a drop of capillary blood is sufficient, then the researcher may abstain from a blood draw by venipuncture.

A second problem is the often large within-person variability which can be of the same size or even larger than between-person variation. This has been observed repeatedly in industrial exposures (see Kromhout et al. 1994), but it is often also true for biological measurements. Liu et al. (1978) reported a ratio of within-person to between-person variances as high as 3.20 for 24-hour urinary sodium, a marker of sodium intake.

The fast development of analytical techniques for biological materials may, however, give rise to many useful indicators. The epidemiologist should be aware of the developments in this area. Methods of which the validity has been assessed should always be preferred, and the manual of operations (see Sect. 13.7.4) should include Good Laboratory Practices (WHO (World Health Organization) 2008). In planning sample logistics and laboratory work, it may be important to keep track of the internal quality control procedures, and provision should be taken that factors influencing sample quality are recorded. The collection of biological samples requires specific provisions and precisely defined procedures to ensure that each processing step is highly standardized, that temperature limits are kept, and that adherence to temporal limits in the processing, shipment, and storage ensures maximum quality of the analytes (Peplies et al. 2011). Especially in multicenter studies, all samples should be analyzed centrally by a certified laboratory to foster standardization and quality assurance. A laboratory information management system (LIMS) may help to document the timing and circumstances of each step of collection, management, transport, and storage for each sample including, for example, temperature logs. An LIMS is a useful tool to trace each aliquot and to document their number, type, volume, and physical location.

Especially in prospective studies, the long-term storage of biological samples is of increasing relevance. On the one hand, the scientific-technological progress allows the identification of new markers and a valid biochemical analysis of continuously shrinking sample volumes. On the other hand, prospective studies allow the targeted analysis of subgroups, for example, in the context of nested case-control studies and thus an efficient and cost-saving approach with maximum gain of knowledge. Such an asservation in a biobank requires the informed consent by study participants.

Measurement of Behavior In recent years, novel technological advances have made it possible to complement the assessment of health-related behavior by the use of monitors. This is particularly useful when the intensity and duration of highly variable behaviors like physical activity have to be measured where even short bouts of activity may be relevant if summed up. For example, accelerometers which are worn for a couple of days are such a device that records any acceleration of the body. The readout of these devices allows to calculate the energy expenditure due to physical activity as well as the pattern and distribution of activities. However, measurements on the subject over a limited time span may be insufficient to cover all of the intra-individual variation of physical activity, and seasonal variations are difficult to assess. Also, modern accelerometers are still limited in that they do not allow the discrimination between different types of activities. Such measurements may therefore require the use of complementary diaries and questionnaires. For further details, refer to chapter ▶Physical Activity Epidemiology of this handbook.

Measurements in the Environment Measurements of physical, chemical, or biological exposures in the environment are best achieved by personal sampling over extended periods of time. Relying on samples collected over relatively short periods of time without clear sampling design may induce substantial error in the exposure assessment. While sensitivity of measurements can be very high, the measurement error caused by the analytical method is often only a small part of the total variance of exposure which is usually dominated by the day-to-day variability in exposure. Planning exposure measurements for an epidemiological study should therefore rely on factors likely to influence the exposure (Sauleau et al. 2003). The methods for exposure measurement have to be included in the manual of operations. Reliable past exposure measurements rarely exist in sufficient numbers. Moreover, their representativeness for the average exposure situation of the study group is doubtful as they were usually obtained for purposes other than an epidemiological study, typically environment control, compliance with TLVs (threshold limit values), or on demand in case of complaints. In such cases an attempt may be needed to assess individual exposure by use of conversion tables such as job-exposures matrices combined with data derived from records, questionnaires, or expert assessment. Moulin et al. (1998a) showed an example where existing past measurements were highly variable and scarcely related to the exposure assessed by experts. Only using the latter, they were able to demonstrate a dose-dependent carcinogenic effect of hard-metal dusts. Nevertheless, even though occupational exposure data are usually not available for each individual study subject, they have been used successfully to model historical exposure levels for specific job tasks and to assign these to individual workers in industry-based cohort studies (e.g., Burstyn et al. 2003) as well as in population-based case-control studies (e.g., Peters et al. 2013).

In other cases, when environmental exposures are monitored continuously as, for example, is the case for close meshed noise registers in the vicinity of airports, the exposure of residents can be assessed with high accuracy, allowing detailed assessment of dose-effect relationships (cf., e.g., the review by Stansfeld and Crombie (2011)).

To assess the impact of the built environment on the behavior of residents such as their physical activity or their food choice, questionnaires can be used to assess the subjective perception of urban forms. In contrast, Geographic Information Systems (GIS) provide geodata on aspects of the built environment to objectively assess urban forms such as street connectivity, land use, or destinations like parks with measures for distance, density, or diversity. Based on appropriate measures of urban forms, indices can be developed to reflect the impact of the built environment as, for example, a moveability index to capture the influence of the built environment on children's physical activity (Buck et al. 2011). Future research will link such environments to actual behaviors by using GPS (global positioning system) monitors.

Finally the use one intends to make of the exposure measurements in the analysis of the data has to be stated in the protocol. This is required in order to assess whether a certain number of measurements with a given precision and a presumed variability of exposure will be sufficient to achieve statistically valid results with respect to the association of exposure and outcome.

13.4.2 Measurement and Classification of Health Outcomes

The main distinction to be made concerning measurement and classification of health outcomes is between diseases for which a clear diagnosis can be made at a precisely defined point in time and health outcomes which are defined by a measurement by either questionnaire or a functional or laboratory test.

In the case of a well-defined disease for which the time of occurrence can be traced individually, both historical cohorts and case-control designs can be used. Still the precise definition of the disease is one of the challenges of a clear design. The main issue is whether to group or to separate disease subtypes. Different subtypes of a same disease may have different risk factors even within a cancer site. The recent rise in adenocarcinoma of the lung has, for instance, been related to the increased use of “light” cigarettes. However, a too narrow definition of a disease may lead to smaller number of cases. This is even more an issue when the diagnosis is obtained from registry data or death certificates as is usually the case in historical cohorts. The precision of such data may be limited, and grouping of diseases thought to have similar etiologies may be a reasonable strategy.

An intermediate case would be a well-defined disease like COPD (chronic obstructive pulmonary disease) for which no systematic recording of patients is available. Here case-control studies face the challenge of estimating the time of onset of the disease in order to ignore all exposures, prior to the occurrence of the disease. Prospective cohorts can circumvent this problem, but if the incidence rate is small, they may not be feasible.

Determination of the point in time when to measure the health outcome is crucial. If one is interested in acute effects of an exposure, for example, the immediate effect of the chlorine in a swimming pool on genotoxicity using the comet assay, the health effect must be measured immediately after the exposure. If one is interested in chronic effects of an exposure, for example, of organic solvents on chronic neurotoxicity, care must be taken to measure the health outcome after a period of washout as, for instance, after the weekend in the case of a study on chronic neurotoxic effects of solvents, as the acute effects of the exposure may be mixed up with its chronic effects. Such considerations may have serious implications on the planning, cost, or even feasibility of such studies.

Often the health outcome of interest is better expressed on a continuous scale rather than classified as diseased/non-diseased like osteoporosis (bone mineral density), chronic obstructive pulmonary disease (FEV1), hyperlipidemia (serum lipid concentration), obesity (body mass index), or hypertension (blood pressure). Then a clear-cut incidence date can hardly be assessed, and the prospective cohort design is best suited to ensure that the exposure precedes the disease.

Similar to what is discussed in the context of exposure assessment, information on health outcomes may be obtained from disease registries as for example, cancer registries that are available in most countries or from claims data where in particular clinical discharge diagnoses can be considered as valid. Use of these individual-level data requires “bullet-proof” procedures for record linkage. Alternatively, information on health outcomes may be obtained from study subjects themselves. However, the ability to recall specific diagnoses is often limited, and underreporting is common, as, for example, in the case of diabetes or hypertension.

13.4.3 Information Bias

A major bias related to data collection is the information bias that occurs if the exposure assessment is different for cases and controls or the health outcome is measured differently for exposed and non-exposed subjects. The first situation arises mainly in case-control studies when cases are worried about the origin of their disease (e.g., smoking or asbestos exposure among lung cancer cases) or deny their disease (an observation made in cancer patients) and therefore recall past exposure differently than controls. It is possible to minimize this type of bias by standardized questionnaires and, if possible, by a data collection blinded with respect to the case-control status. The latter approach, although an option in hospital-based case-control studies, is, however, virtually impossible in population-based case-control studies. The same type of information bias is possible with historical cohorts if information is obtained a posteriori from proxies.

An information bias can also be due to the unavoidable measurement error in retrospective exposure assessment. The comparable accuracy principle (Wacholder et al. 1992a) states that the degree of accuracy in measuring the exposure of interest for the cases should be equivalent to the degree of accuracy for the controls unless the effect of the inaccuracy can be controlled in the analysis, as, for instance, by using appropriate validation data. Although adherence to this principle does not eliminate the corresponding misclassification, its rationale is to avoid that a positive finding is induced simply by differences in the accuracy of information about cases and controls (see also chapter ▶Case-Control Studies of this handbook).

13.5 Confounding

The *deconfounding principle*, another principle given in Wacholder et al. (1992a), states that confounding should not be allowed to distort the estimation of an effect. Confounders are by definition factors that are determinants of the disease and that are related to the exposure of interest (see also chapter ▶Confounding and Interaction of this handbook). Setting up a list of confounders is therefore a primary task in the exploration of the scientific knowledge related to the research question of the planned study. This identification of confounders should not be based on statistical associations alone. It must also include knowledge of potential causal pathways. For instance, if one were to study the effect of nutrition on coronary heart diseases (CHD), a factor that might be considered a potential confounder is obesity as it is related to both nutrition and CHD. But one might consider that obesity is on the causal pathway between nutrition and CHD in which case controlling for obesity would bias the estimation of the effect of nutrition (*overadjustment*). Careful thought should therefore be given to each factor that has to be included as a confounder. An operational result of this first step would be a list of three groups of confounders:

(i) Established confounders that are known to determine disease and to be related to the exposure of interest but which are not on the causal pathways. Sex and age are virtually always included in this group. Such confounders are to be included even if they do not show a statistically significant association with the disease in the study sample.

(ii) Probable confounders that are established risk factors and for which there are reasons to believe that they might be related to the exposure.

(iii) Possible confounders including other risk factors of the disease that might be or not related to the exposure of interest. If a given risk factor is a strong determinant of the disease (e.g., smoking for lung cancer), it might confound the association of a potential risk factor although the confounder is only weakly associated with the exposure of interest in the study sample.

At the design stage, controlling for confounders can be done either by restricting the study sample to certain values of the confounder (e.g., a lung cancer case-control study among non-smokers) or by matching. The latter ranges from broad frequency matching on age and sex to the individual matching on factors thought to be related to non-measurable factors (e.g., neighborhood matching). Matching aims at equalizing the variation of the confounder between cases and controls. This should, however, not reduce the differences and the variation of the exposure of interest between both groups. Nevertheless, if the matching variable could also serve as a proxy for the exposure of interest, it may happen that the effect of the exposure is decreased or even eliminated by matching.

This so-called overmatching is one of the main pitfalls in controlling for confounders at the design stage. It must also be stressed at this point that controlling for confounders in the design may well be counterproductive as it is irreversible and may forbid to explore interesting post hoc hypotheses (see also Sect. 13.3 on matching in chapter ▸Case-Control Studies of this handbook). A last strategy is to collect the relevant data on confounders in order to be able to control for them later in the analysis. Details of how to deal with confounding using any of these methods can only be decided within the context of a given study.

13.6 Statistical Analysis

Enough time and human resources should be planned from the start of a study for the statistical analysis. Many large-scale studies we know of are underreported because of lack of resources for the statistical analysis. It is impossible to detail all statistical analyses already in the planning stage. Nevertheless, the main research questions should be formulated in the protocol, and these questions should be operationalized

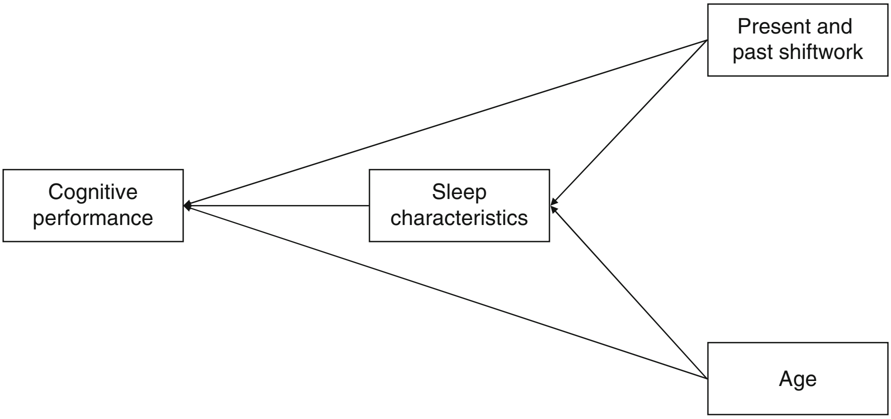


Fig. 13.3 Causal hypotheses of effects of shiftwork and age on cognitive performance

already at this stage, that is, translated in a statistical hypothesis to be tested. In addition, a statistical analysis plan should be developed at least for analyzing the main research question. General principles as well as special applications of statistical analyses of epidemiological studies are described in Part III (Statistical Methods in Epidemiology) of this handbook.

It is helpful at this stage to draw hypothetical causal graphs describing a priori models to be applied to the data of the study (see Greenland et al. 1999; Hernán and Robins 2006; Pearl 1995, 2000; chapter ▶Directed Acyclic Graphs of this handbook). Figure 13.3 shows a very simple graph formulating a series of hypotheses on the effect of shiftwork and age on cognitive performance which may be mediated by sleep problems. Such a conceptual framework helps to develop a strategy for statistical analyses based on a priori hypotheses and to identify potential confounders.

Additional ad hoc analyses which are more or less data driven may follow. The minimum statistical contents to be defined at the planning stage are the endpoint variable(s), the exposure variables, and the confounder variables to be adjusted for. The statistical measures (proportions, odds ratios, means, standard deviations, etc.) and models (mostly regression models) to be calculated should also be fixed in advance. Moreover, sensitivity analyses should be foreseen to check the robustness of the results against deviations from model assumptions. Finally, the subgroups for which stratified analyses will be done should be identified.

The fact that the a priori hypotheses are operationalized means that, for instance, the main endpoint(s) is clearly defined by certain values of the outcome variable(s). This is relatively straightforward in case-control studies for which the endpoint is the case status whatever its definition. For cohort studies investigating the mortality or incidence of diseases, the list of diseases possibly related to the exposure of interest should be well defined. For prospective cohorts or cross-sectional studies, the endpoint of interest may be less straightforward and can possibly require data transformation, for example, a score of depression obtained from a mental health questionnaire.

The variables characterizing the exposure should also be precisely specified. These can include (possibly lagged) cumulative, peak, or mean exposure metrics. Although for power computations this exposure may have to be considered as a yes/no variable, the main message cannot be that simple. Established confounders have to be included in the final model in any case. However, although the robustness of an association should be tested against all measured confounders, the list of presumed confounders to be included in the main model cannot be finalized at the planning stage.

13.7 Practical Issues

13.7.1 Fund Raising and Budgeting

Usually, large-scale epidemiological studies require external funds which can be acquired from industry or from public sources like foundations, research organizations, or governmental research funding. Often, the latter are offered through thematic calls where competing proposals are submitted and then evaluated by an ad hoc board of scientific experts as is the case in the periodic Framework Programmes of the European Commission. Sometimes the structure according to which such project proposals are to be set up is given by the funding agency. In general, the points to be addressed overlap to a large degree with what is required for the study protocol (see Box 13.1) where a detailed cost plan has to be added and the major cost items have to be justified. Technically, the expenses have to be broken down as regular salaries (paid by the parent organization), salaries for temporary staff, durable equipment, travel expenses, consumables (mailing, telecommunication, office material, laboratory), fees, consultants, subcontracts, overhead costs, and, if applicable, VAT.

It is also important to adjust the disbursement of the budget to fit the respective stages of the project. If funding comes from several different sources, the tentative share of each one must be specified. Experience has shown that the situation is easier to handle if only one funding agency is involved, but very large projects may require more than one source of funding (Hernberg 1992, p. 189f).

Once a grant application is successful, regardless of whether the sponsor is private or public, a contract should be set up between the researcher and the sponsor that:

(a) Ensures independence of the research project by prohibiting interference of the sponsor with the study conduct, analysis, and interpretation
(b) Guarantees full publication rights to the researcher, that is, that the results will be submitted for publication regardless of outcome and without a veto of the sponsor
(c) Defines the deliverables, the timeline, as well as the budget and its installment plan
(d) Regulates the intellectual property rights

Regarding the latter, the IEA (International Epidemiological Association) states in its guidelines for proper conduct of epidemiological research (Good Epidemiological Practice (GEP), IEA 2007) that “All results of a study, whether government or industry-sponsored, should be the intellectual property of the investigators, not the sponsor. Requests to withhold findings, to change or tone down the content of a report to produce a misleading or delayed publication should be categorically refused.”

13.7.2 Data Management

Data management is a crucial step especially with multicenter studies (see Long (2009) for a recent account on these issues, chapter ▸Data Management in Epidemiology of this handbook).

A first task in planning the data management is to identify the different data tables which are to be set up and to plan their structure and their linkage. Usually, data protection requires that all direct identifiers like name and address or social security number have to be kept separately from data tables (pseudonymization) or even that direct identifiers have to be deleted immediately after data collection (anonymization). In both cases a numerical subject identifier is assigned to each study subject and added to the corresponding data records which in itself does not allow the re-identification of the individual.

For linkage with external data, be it an administrative file or a file containing causes of death or diseases, as it is often necessary in (historical) cohort studies, a data table must be set up containing only the minimal information for linkage. Usually this entails the pseudonymized number used in the analysis dataset and the information by which linkage is facilitated, for example, name and address or a unique personal identifier like the social security number. The table used for linkage is provided to the institution holding the external data. External records matching with the direct identifiers can then be sent back without the direct identifiers and merged with the study data by the numerical subject identifier.

The data management of prospective cohorts and other population-based studies is especially challenging as there are several dimensions in the data. It is nearly unavoidable in such studies to keep a separate file for the management of all contacts. In this file all letters sent and received, all telephone contacts, all datasets, or information received should be documented for every study subject. Such a (complex) recruitment system allows to quickly identify non-responders, to control adherence to the recruitment protocol and to follow up the study subjects. If a subject has moved, for instance, such a system can support tracing and recording of new addresses. It may also ensure that no repeated letters are sent to deceased subjects. It is important to know at each moment if a given subject is due for another contact and whether she/he has reacted to the last mailing or phone call. This also implies that a recruitment system is able to trigger mailing of the correct letter according to the subject's status (questionnaire to be sent, questionnaire received, lost to follow-up, pending questionnaire, tracing of new address initiated, etc.). The contact data received must always be kept separately from the study data.

It is also important to clearly identify and label the variables containing the same information collected at different points in time when planning a longitudinal study. If these variables have the same names, they are at risk to be overwritten. Careful a priori structuring of the database makes life much easier when creating a dataset for statistical analysis. Owing to the extended time scale of such studies, the documentation is particularly important as it is not guaranteed that the same data managers will handle the data throughout the study.

Other issues to be planned carefully are data entry and coding. Three main options exist for data entry. A first option is to input data directly when interviewing the study participants. This is the rule with telephone interviews, and it is widely used with laptops in face-to-face interviews. An important aspect with direct input is the layout of the screen; the interviewer must follow the questions she/he has to ask without being disturbed by computer problems. Issues like skip patterns, the ability to easily correct already entered data, toggling between keyboard and mouse, and online detection of invalid codes or inconsistencies between different data items are to be programmed carefully and to be tested under real conditions. A widely used computer program is the freeware Epi Info available from the division of public health surveillance and informatics of the Centers of Disease Control, Atlanta (Centers for Disease Control and Prevention 2011). However, it has some limitations making it unsuitable for complex large-scale studies. A large range of commercial software exists, each with its own advantages (see chapter ▶Data Management in Epidemiology of this handbook). A second option is to obtain data through paper questionnaires and to enter the data later. In this situation a double entry is to be recommended whenever possible, especially if the questionnaires were self-administered. The data should be entered as they are, but the data problems (errors, inconsistencies) should be documented in a log-file to be cleared as soon as possible. Another option is to obtain machine-readable questionnaires which are scanned electronically. Such an approach has been implemented in actual large-scale studies, notably in the French part of the EPIC study (e.g., Clavel-Chapelon et al. 2002). Future developments will make the use of electronic self-completion instruments more accessible, in particular in combination with user-friendly touch screens and online access allowing real-time data transfer to a secure central data server (see chapter ▶Internet-Based Epidemiology). Such online tools can be easily accessed from anywhere as they only require Internet access and a usual web browser (Lumsden 2005; Reynolds et al. 2007).

Data coding, that is, transforming free text information (examples are places of residence, jobs, tasks, or food items), in a closed list of items is also an aspect to be planned and tested. It relies often on specialized knowledge. The closed list of items and coding rules must be set up before starting the data collection. Details on the construction of instruments are given in chapter ▶Epidemiological Field Work in Population-Based Studies of this handbook.

The management of stationary exposure data deserves a specific discussion since they pertain to exposure groups rather than individuals, that is, they characterize specific circumstances (cf. chapter ▶Exposure Assessment of this handbook). These circumstances include, for example, measurements of air pollution in certain areas, households, or occupational tasks. These data must entail a link to the subject data. It is therefore important to include the same items (i.e., labels of the exposure group) in the exposure measurement database as in the subjects' database. A further complication arises when some of these exposure measurements are on individuals (e.g., personal sampling at the workplace or stationary measurements in households), including subjects from the epidemiological database. These measurements characterize both the exposure group and the specific individual. Both links must then be clearly identified from the start.

Finally, all steps of data management should be recorded in a logbook. This can be part of the overall project diary (see below).

13.7.3 Quality Assurance

Industrial standards for quality assurance and quality management are set down in the ISO 9000 series of standards (International Organization for Standardization 2012). Their application to epidemiological studies is not straightforward given that these standards (see Moulin et al. 1998b) are geared toward customer satisfaction and that the customers of epidemiology are not easily defined. The main ideas behind these standards are however useful.

The main principles as they apply to epidemiology are the following (see also chapter ▶Quality Control and Good Epidemiological Practice of this handbook). Write up in detail what you intend to do and document what you did. Try to be proactive in thinking of what can go wrong and plan accordingly. Set up means by which you can detect any problem as early as possible and by which you can correct your procedures accordingly. Specify and document all changes of procedures implemented after the start of the study.

We already insisted on the necessity of a detailed study protocol. A study protocol may be complemented by a manual of operations which includes all standard operating procedures (see below) describing the actual work to do (cookbook). One main point in being proactive is to specify all procedures for data collection in as much detail as possible. Details with respect to material conditions, for example, hardware, software, office, and storage room, need to be considered in advance. Training of the data collection staff is a key to good quality data. This training should encompass all survey instruments in the setting in which the actual data collection will be conducted. It is also important to acknowledge that there will be non-responders and to collect minimal information on non-participants from the beginning to assess the type and degree of a potential selection bias. In order to monitor errors as they arise and to enable immediate corrective action, data should be entered concurrently to their collection, and data checks and validation should be done as early as possible. In large-scale studies, the internal quality management may be supported by an independent external quality control (for an example, see Filipiak-Pittroff and Wölke 2007).

13.7.4 Study Conduct: Manual of Operations

As mentioned in the section on the study protocol, the collection of material and data as well as the methods and procedures must be described in detail. This can be done in a separate document: the manual of operations (cf. chapter ▸Quality Control and Good Epidemiological Practice of this handbook). The manual of operations compiles all standard operating procedures. It may therefore be considered as the technical operationalization of the study protocol which serves as the "cookbook" for the study personnel and ensures standardization of all measurement procedures. By this it becomes a central part of the quality management of a study. An exemplary table of contents of a manual of operations is given in Box 13.2.

Box 13.2. Table of contents of an exemplary manual of operations taken from the IDEFICS study (Ahrens et al. 2011)

1 General
1.1 Background and research question
1.2 Description of study base and study region
1.3 Inclusion and exclusion criteria for study subjects
1.4 Sampling design including instructions for drawing the sample
1.5 Public relations and corporate design
2 Fieldwork
2.1 Contact procedures and approaching study subjects
2.2 Informed consent
2.3 Treatment of participants and non-participants
2.4 Electronic control and documentation of contacts
2.5 Assignment of ID-numbers (pseudonymization)
2.6 Procedures for data protection
3 Data collection
3.1 Overview of data collection modules
3.2 Organization of data collection
4 Interviews and questionnaires
4.1 Structure and contents of questionnaires
4.2 Instructions for use of questionnaires/conduct of interviews

4.3 Editing of completed questionnaires
4.3.1 General instructions
4.3.2 Specific instructions
(a) parental questionnaire, (b) dietary assessment, (c) medical history, (d) teachers questionnaire, (e) setting questionnaires (schools and kindergartens)

5 Physical examinations
5.1 General rules and procedures
5.1.1 Functional checks of measurement devices
5.1.2 Completion of documentation sheets
5.1.3 Instructions for clothing of children during examinations
5.1.4 Hygiene
5.2 Anthropometric measurements
(a) body weight and bio-impedance, (b) body height, (c) waist and hip circumference, (d) skinfold thickness, (e) . . .

5.3 Blood pressure and heart rate
5.4 Actimetry (accelerometer)
5.4.1 Preparation of measurement devices
5.4.2 Measurement protocol and implementation of measurement devices
5.4.3 Readout and taking readings
5.4.4 . . .
6 Biological samples
6.1 General procedure for collection, processing, storage and transport
6.2 Specific instructions
6.2.1 Blood samples
6.2.2 Urinary samples
6.2.3 Salivary samples
6.2.4 . . .
7 Handling of devices
7.1 Scale with bioelectrical impedance measurement
7.2 Stadiometer
7.3 Skinfold calliper
7.4 Blood pressure measurement
7.5 Accelerometer
7.6 . . .
8 Quality management
8.1 Training of study personnel
8.2 Pretest of questionnaires and measurement procedures
8.3 Re-interviews and repeated measurements
8.4 Monitoring of data collection and site visits
8.5 External quality control
9 Ethical principles and data protection

10 Annex
10.1 Questionnaires and documentation sheets
10.2 Letter templates
10.3 Information material and consent declaration
10.4 Important contacts
10.5 Checklist supporting adherence to standard operating procedures (SOPs)
10.6 Checklist for regular functional checks of measurement devices
10.7 Reporting form for daily inspection of the equipment of the examination site
10.8 List of all devices and consumables
10.9 Instructions for parents

It describes in detail the recruitment of study subjects and how they are selected for the study. The eligibility criteria must be defined in cohort studies in terms of minimum exposure, calendar time of exposure, whether or not other exposures are allowed, and so forth. If a case-control design is adopted, all eligibility criteria except the diagnosis for both cases and controls must be comparable to ensure that both groups are drawn from the same study base. For example, what histological type of cancer will be included and how will the diagnosis be confirmed? What is the data source for identification of study subjects? Who will collect the data? What is the sequence of approaches to contact a study subject (letter, phone call, home visit), and what are the criteria to classify a subject as a non-responder?

The measurement methods and procedures need to be described as well. How will the interviewers be trained? Will there be a panel of pathologists for reviewing the diagnosis based on the histological slides? How are the questionnaires to be applied to study subjects? What are the exact steps and procedures in the measurement of blood pressure (posture of the study subject, resting time before taking the first measurement, number of measurements, time interval between subsequent recordings, choice of the correct cuff size, advice given to the respondent, etc.)? What is the definition of fasting, for example, what is the minimum time interval between the last meal and the blood draw of the subject? In the case of biological samples, the manual needs to describe exactly all procedures and the timing of collection, processing, shipment, and storage of samples to avoid biased measurements due to decay of metabolites. It may also include a description of the analytical laboratory procedures.

Moreover, the manual of operations should summarize the ethical principles of the study, the information to be given to study participants, and the procedures to ensure confidentiality during data collection and data management. It should entail all information that is relevant for the daily routine of the study personnel during fieldwork and data management.

13.7.5 Timeline and Workflow

The time schedule concerns the sequence and interrelationship of different operational tasks and resources. It is good practice to outline the tasks and subtasks at the design stage and to plan their time flow which can be visualized in a Gantt chart (after the method developed by Charles Gantt in 1917; for a simple fictitious example, see Fig. 13.4). A PERT (Program Evaluation and Review Technique) chart (for a fictitious example, see Fig. 13.5) gives a quick overview of the relationships between main tasks and helps to identify the critical paths. This will help to define milestones that need to be attained in time in order to successfully reach the aims of the project. It should be noted that milestones are not to be mixed up with deliverables. The latter define the end result of a given activity or task, while the former define a more complex stage that has to be reached in the overall roll-out of a project in order to enter the subsequent phase. Responsibilities have to be assigned for each task, and the necessary resources required for each have to be allocated. Once it has been decided how many subjects will be included and what methods of examination will be used, the time needed can be estimated rather accurately. At the planning stage, one should make sure that statistical and computer support will be available when needed. One should realize that the first data analysis may reveal unforeseen inconsistencies or other problems in the data that then trigger subsequent analyses not foreseen at the beginning.

Writing a manuscript takes time. Unexpected practical matters almost always disrupt the original time schedule. Enough time must be reserved for all these considerations. Hernberg (1992) recommended to foresee an allowance of half a year or more for unexpected complications.

The timeline and the corresponding resources are best planned using project management tools like Gantt and PERT charts (see Figs. 13.4 and 13.5). Basically these tools decompose a project in elementary tasks with certain duration, order, and interrelationship. Figure 13.5 presents the different tasks with arrows indicating which tasks must be terminated before the next one can start (e.g., the data collection

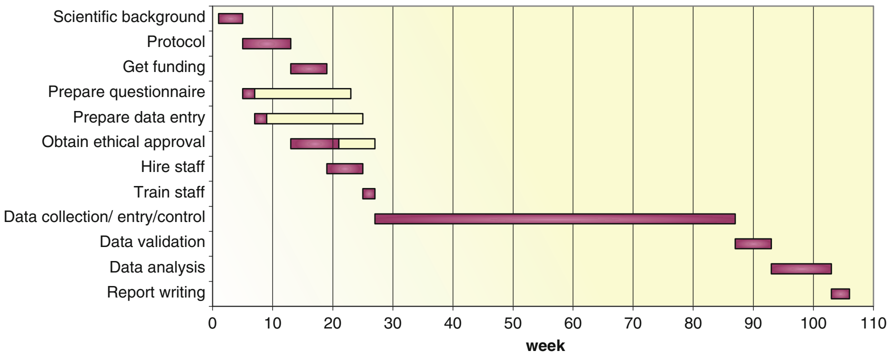


Fig. 13.4 Gantt chart for planning a study

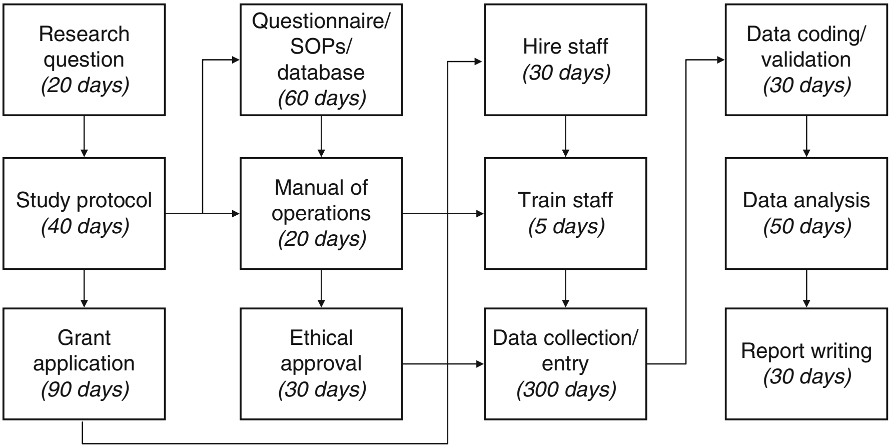

Research question (20 days)
Questionnaire/ SOPs/ database (60 days)
Hire staff (30 days)
Data coding/ validation (30 days)
Study protocol (40 days)
Manual of operations (20 days)
Train staff (5 days)
Data analysis (50 days)
Grant application (90 days)
Ethical approval (30 days)
Data collection/ entry (300 days)
Report writing (30 days)

Fig. 13.5 Simple fictitious example of a PERT chart for planning a study

can only start when the study has been approved by the ethics committee and when the staff has been trained). From this information, critical tasks (dark bars in Fig. 13.4) are identified, which – if delayed – will delay the whole study. Non-critical tasks such as preparation of data entry that may depend on completion of other tasks as, for instance, preparation of questionnaire can be delayed without endangering the successful attainment of milestones. However, non-critical tasks must be completed by the start of other activities (here, data entry). This is indicated by the light bars in Fig. 13.4. For examples of project planning tools and an introduction to project management in general, refer, for instance, to Schwalbe (2012). A number of sharewares easily available through the Internet provide the software to draw these charts and diagrams.

13.7.6 Project Diary

A project diary should trace the major activities performed during the course of the project. This includes the changes of the protocol, the data collection, and the data processing. A diary keeps track not only of the scientific realization of the operational plan of the project but also of its administrative and economic aspects.

Changes of the original work plan may be dictated either by external circumstances possibly by scientific reasons, for example, if new evidence arises after the start of the study concerning a potentially important confounder.

Documentation of data collection should monitor not only its advancement but all potentially important events that might distort measurement values. It should thus be documented if personnel involved in collection or management of data is replaced, if some measurement instrument is replaced or if and why, at some stage, the data collection is impaired. Temperature loggers facilitate documentation of involuntary thawing of frozen biosamples. Dates and results of instrument calibration should be recorded. When quality control measures result in corrective action (see chapter ▶Quality Control and Good Epidemiological Practice), this should also be recorded.

Electronic data handling, for example, names of raw data files, computer transfers, plausibility checks, data corrections, merging of files, recoding, and finally all statistical data processing, should all be documented including analyses which proved to be dead ends (cf. chapter ▶Data Management in Epidemiology of this handbook).

The project diary and the protocol, operation manual, and the raw data are the central pieces of evidence with which errors can be traced and the study can be replicated or be reanalyzed, if needed.

13.7.7 Ethical Aspects

The ethical aspects are diverse and have been well covered in chapter ▶Ethical Aspects of Epidemiological Research of this handbook. General ethical principles for all biomedical research have been laid down in the Helsinki declaration of the World Medical Organization (1996). The four central principles can be summarized as follows:

1. Doing good (beneficence)
2. Not harming (non-maleficence)
3. Respecting persons (autonomy)
4. Distributing goods and evils fairly (justice).

The following aspects are more specific to epidemiological research: The first ethical requirement is the respect of the data protection laws that are specific to each country. This may even require deletion or at least separation of parts of the data, usually the personal identifiers, at a certain point in time. The second is the requirement that all study participants are informed of the objectives of the study, which medical examinations will be done and what is their purpose and associated health risks. The written informed consent of the participants is obligatory. Confidentiality must always be adhered to, and no individual results can be revealed to third parties unless explicitly permitted by the respondent. If in the course of a prospective study it becomes evident that individual exposure levels exceed safety limits, the researcher must take the initiative to try to remove the subject from the hazardous exposure.

The results of individual investigations should be made available to study participants upon request as well as a summary of the overall results of the study. However, incidental findings pose a special problem. A medical examination might detect a hidden disease, for example, an X-ray or an MRI might detect a tumor. Participants have the right to know such findings, but they may also dissent to get informed when giving their informed consent. If the result of an examination is of therapeutic or preventive relevance, the study participant may be referred to a physician. Otherwise, pathological findings that are not amenable to treatment or prevention may be excluded from feedback.

Biological markers have become a major issue of concern with regard to the right to know or not to know mentioned above. Here, in planning a study, three types of markers should be distinguished. The first group is formed by biological characteristics with yet unknown relevance for the risk or prognosis of a disease that are under research and for which no reference values exist. This is the case for numerous genetic markers with unknown biological function. In this case knowledge of this value does not entail a direct benefit for study participants. The second group of markers is those that might have a direct therapeutic implication as in the case of blood lipids or certain chemical noxae. In this case it may be an ethical obligation to inform the affected study participant in due time about elevated or pathological values to avert any damage. The third group includes markers that are predictive of future disease but without any direct implications for therapy or prevention as in the case of some genetic disorders. In this case, the right to know or not to know is of particular importance. In such cases it should be carefully evaluated, whether the assessment of such markers is really necessary to reach the study aim. If this is the case, it may be an option to inform study participants that such information will not be disclosed. It is recommended to consult the responsible institutional review board/ethics committee how to proceed.

A final aspect is that the results of the study should always be made available to the community. Failing to publish the results means that the examinees have been abused and the funding has been wasted. In doing so, the interpretation of the results must be objective including a discussion of all relevant literature and all possible validity problems as well as alternative explanations of the findings. Further details and specific recommendations relevant for epidemiological studies were formulated by the Council for International Organizations of Medical Sciences in collaboration with the WHO (CIOMS 2009).

13.7.8 Scientific Collaborations and Multicenter Studies

An epidemiological study usually requires collaborations of the principal investigator with scientists from several fields which of course include epidemiology and usually a specialized medical field and statistics but may also include, depending on the study, genetics, molecular biology, microbiology, chemistry, industrial hygiene, psychology, and sociology. Another type of collaboration occurs if a multicenter study is set up in which several epidemiologists combine resources. Advantages and possible pitfalls have already been mentioned, and the standardization of the data collection is then a major issue. The data analysis may be centralized or decentralized. For instance, in multicenter studies organized by the International Agency for Research on Cancer (IARC), the national centers may publish their local data, but they may also take the lead for a pre-defined specific analysis of a particular topic of the pooled dataset. Usually, publications based on the pooled full dataset have priority over local analyses.

In any type of collaboration, the respective responsibility and role of each collaborator or collaborating center as well as the resources allocated to each partner should be settled before the study starts. These are the key components of the project governance. Provisions should be made for possibly divergent interests between study partners by laying down all important rights and obligations of each partner in a contract. This must include the regulation of intellectual property rights as well as access rights to data and results. Responsibilities for deliverables and corresponding deadlines should become part of such a contract, for example, by appending the study protocol. Establishment of a democratic governance structure has proven to be useful in large-scale multicenter studies, for example, by establishing a general assembly of all principal investigators. A steering committee consisting of the leaders of major work packages may support the risk management by initiating actions needed to follow the study plan and by adapting the plan if necessary. It may also take care that all partners collaborate on an equal basis (taking their contribution to the project into account) and get their share of the scientific output including publications.

13.7.9 Communication

Epidemiological research is of interest not only to epidemiologists but also to decision makers, funding agencies, and also the general public. The results should be published in a form and language that they are understandable to the different target groups. This often requires two or more levels of reporting, one scientific and one popular, possibly supported by a press release. It is worthwhile to plan responsibilities for each aspect in advance and how the information will be disseminated. Those taking part in a study have the right to know not only their personal results – especially when medical examinations are involved – but also, at least in general terms, the outcome of the whole study. The correct timing of the sequence of information delivery is important. First, those examined should be informed of their own results unless a subject declared that he/she does not want to be informed; then summary results should be given to funding agencies and only afterward to the news media. Ideally, a peer-reviewed scientific article should have appeared or at least have been accepted for publication before informing the news media. However, the scientific publishing procedure is often slow that it sometimes may be unethical to withhold urgent results from the public that long.

According to the precautionary principle, epidemiologists are responsible to communicate a justified suspicion regarding potential health hazards even if the evidence of the associated risk is still limited. In doing so, an epidemiologist has to balance out the information on potential risks and their implications and the communication of limitations and uncertainties of the study findings. This has been pointed out by Sandman (1991) in his guidelines on communication responsibilities of epidemiologists regarding health risks: “I do argue that poor communication comprises even the best epidemiology, and that epidemiologists therefore have communication responsibilities that cannot be ignored.”

A large project usually gives rise to several scientific publications, and it may be useful to outline their topics in advance. At the planning stage, it is advisable to agree within the team on who will be responsible, that is, the first author, and of what and how authorship will be facilitated. It is usually not possible to decide the order of names at this stage because each team member's input to the intellectual process can be judged only after the corresponding activity has been successfully completed. Publication rules may become part of the contract between researchers involved in a study as, for example, in the IDEFICS study (2011). Central guidelines as to who should be considered author of a publication are included in the Vancouver guidelines (International Committee of Medical Journal Editors (ICMJE) (2010)). These include the obligation to declare any potential conflicts of interest to avert doubts about the independence of the research.

13.8 Conclusions

Careful planning is a key in the successful completion of an epidemiological study. This planning should be based on up-to-date scientific knowledge and awareness of all possible pitfalls inherent in epidemiological studies. It should cover all aspects from study base definition, precise design used, statistical power, control of confounding, precise data collection, and exposure measurement methods to quality control, statistical methods, collaborations, dissemination of study results, and ethical issues. All these issues should be laid down in a scientific study protocol, and a manual of operations from which the quality of the study can be assessed. Multicenter studies require regulations to ensure respectful and fair collaboration between all partners. Finally, researchers have to admit their responsibility to communicate their findings not only to the scientific community but also to the public.

References

Ahrens W (1999) Retrospective assessment of occupational exposure in case-control studies. Development, evaluation and comparison of different methods. Ecomed, Landsberg

Ahrens W, Bammann K, Siani A, Buchecker K, De Henauw S, Iacoviello L, Hebestreit A, Krogh V, Lissner L, Mårild S, Molnár D, Moreno LA, Pitsiladis YP, Reisch L, Tornaritis M, Veidebaum T, Pigeot I, IDEFICS Consortium (2011) The IDEFICS cohort: design, characteristics and participation in the baseline survey. Int J Obes 35(Suppl 1):S3–S15

Ahrens W, Jöckel KH, Brochard P, Bolm-Audorff U, Grossgarten K, Iwatsubo Y, Orlowski E, Pohlabeln H, Berrino F (1993) Retrospective assessment of asbestos exposure–I. Case-control analysis in a study of lung cancer: efficiency of job-specific questionnaires and job exposure matrices. Int J Epidemiol 22(Suppl 2):S83–S95

Andersen LF, Bere E, Kolbjornsen N, Klepp KI (2004) Validity and reproducibility of self-reported intake of fruit and vegetable among 6th graders. Eur J Clin Nutr 58:771–777

Armstrong BK, White E, Saracci R (1994) Principles of exposure measurement in epidemiology. Oxford University Press, Lyon

Biobank UK (2012) http://www.ukbiobank.ac.uk. Accessed 3 Sept 2012

Boeing H, Korfmann A, Bergmann MM (1999) Recruitment procedures of EPIC-Germany. European Investigation into Cancer and Nutrition. Ann Nutr Metab 43:205–215

Boffetta P, Burstyn I, Partanen T, Kromhout H, Svane O, Langård S, Järvholm B, Frentzel-Beyme R, Kauppinen T, Stücker I, Shaham J, Heederik D, Ahrens W, Bergdahl IA, Cenée S, Ferro G, Heikkilä P, Hooiveld M, Johansen C, Randem BG, Schill W (2003) Cancer mortality among European asphalt workers: an international epidemiological study. II. Exposure to bitumen fume and other agents. Am J Ind Med 43(1):28–39

Bogers RP, Van Assema P, Kester AD, Westerterp KR, Dagnelie PC (2004) Reproducibility, validity, and responsiveness to change of a short questionnaire for measuring fruit and vegetable intake. Am J Epidemiol 159:900–909

Breslow NE, Chatterjee N (1999) Design and analysis of two-phase studies with binary outcomes applied to Wilms tumor prognosis. Appl Stat 48:457–468

Breslow NE, Day NE (1981) Statistical methods in cancer research, vol I-The analysis of case-control studies. IARC, Lyon

Breslow NE, Day NE (1987) Statistical methods in cancer research, vol II-The design and analysis of cohort studies. IARC, Lyon

Buck C, Pohlabeln H, Huybrechts I, De Bourdeaudhuij I, Pitsiladis Y, Reisch L, Pigeot I (2011) Development and application of a moveability index to quantify possibilities for physical activity in the built environment of children. Health Place 17(6):1191–1201

Burstyn I, Boffetta P, Kauppinen T, Heikkilä P, Svane O, Partanen T, Stücker I, Frentzel-Beyme R, Ahrens W, Merzenich H, Heederik D, Hooiveld M, Brunekreef B, Langård S, Randem BG, Järvholm B, Bergdahl IA, Shaham J, Ferro G, Kromhout H (2003) Performance of different exposure assessment approaches in a study of bitumen fume exposure and lung cancer mortality. Am J Ind Med 43(1):40–48

Centers for Disease Control and Prevention (2010) National Health and Nutrition Examination Survey NHANES I. http://www.cdc.gov/nchs/nhanes/nhefs/nhefs.htm. Accessed 7 Sept 2012

Centers for Disease Control and Prevention (2011) Epi Info^{TM}. http://www.cdc.gov/epiinfo. Accessed 26 July 2012

CIOMS (2009) International ethical guidelines for epidemiological studies. Council for International Organizations of Medical Sciences (CIOMS) in collaboration with the World Health Organization (WHO), Geneva

Clavel-Chapelon F, E3N-EPIC Group (2002) Differential effects of reproductive factors on the risk of pre- and postmenopausal breast cancer. Results from a large cohort of French women. Br J Cancer 86:723–727

Filipiak-Pittroff B, Wölke G (2007) External quality assurance in the German Health Interview and Examination Survey for Children and Adolescents (KiGGS). Procedure and results (in German). Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz 50(5–6): 573–577

Flesch-Janys D, Steindorf K, Gurn P, Becher H (1998) Estimation of the cumulated exposure to polychlorinated dibenzo-p-dioxins/furans and standardized mortality ratio analysis of cancer mortality by dose in an occupationally exposed cohort. Environ Health Perspect 106(Suppl 2):655–662

Garfinkel L (1985) Selection, follow-up, and analysis in the American Cancer Society prospective studies. NIH Publication 85–2713, National Cancer Institute Monograph 67. Government printing office, Washington, DC, pp 49–52

Goebel A, Netal S, Schedel R, Sprotte G (2002) Human pooled immunoglobulin in the treatment of chronic pain syndromes. Pain Med 3:119–127

Goldberg M, Chastang J-F, Leclerc A, Zins M, Bonenfant S, Bugel I, Kaniewski N, Schmaus A, Niedhammer I, Piciotti M, Chevalier A, Godard C, Imbernon E (2001) Socioeconomic, demographic, occupational, and health factors associated with participation in a long-term epidemiologic survey: a prospective study of the French GAZEL cohort and its target population. Am J Epidemiol 154:373–384

Gordis L (2009) Epidemiology, 4th edn. Saunders-Elsevier, Philadelphia

Greenland S, Robins J (1994) Invited commentary: ecologic studies-biases, misconceptions, and counterexamples. Am J Epidemiol 139:747–760

Greenland S, Pearl J, Robins JM (1999) Causal diagrams for epidemiologic research. Epidemiology 10:37–48

Hernán MA, Robins JM (2006) Instruments for causal inference: an epidemiologist's dream? Epidemiology 17:360–372

Hernberg S (1992) Introduction to occupational epidemiology. Lewis Publishers, Chelsea

IARC (International Agency for Research on Cancer) EPIC Project (2012) http://epic.iarc.fr. Accessed 23 Jan 2012

IDEFICS study (2011) IDEFICS publication rules. http://www.ideficsstudy.eu/Idefics/UserFiles/File/Pubrules_IDEFICS_ver_2011Jun.pdf. Accessed 30 Sept 2012

IEA (International Epidemiological Association) (2007) Good epidemiological practice (GEP): IEA guidelines for proper conduct of epidemiologic research. http://ieaweb.org/2010/04/good-epidemiological-practice-gep/. Accessed 29 Sept 2012

International Committee of Medical Journal Editors (ICMJE) (2010) Uniform requirements for manuscripts submitted to biomedical journals: writing and editing for biomedical publication. http://www.icmje.org/urm_full.pdf. Accessed 30 Sept 2012

International Organization for Standardization (2012) ISO 9000 – Quality management. http://www.iso.org/iso/home/standards/management-standards/iso_9000.htm. Accessed 26 July 2012

Kaslow RA, Ostrow DG, Detels R, Phair JP, Polk BF, Rinaldo CR Jr (1987) The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. Am J Epidemiol 126:310–318

Kreienbrock L, Pigeot I, Ahrens W (2012) Epidemiologische Methoden, 5th edn. Springer, Berlin/Heidelberg

Kromhout H, Symanski E, Rappaport SM (1994) A comprehensive evaluation of within- and between-worker components of occupational exposure to chemical agents. Ann Occup Hyg 37:253–270

Lagiou P, Georgila C, Minaki P, Ahrens W, Pohlabeln H, Benhamou S, Bouchardy C, Slamova A, Schejbalova M, Merletti F, Richiardi L, Kjaerheim K, Agudo A, Castellsague X, Macfarlane TV, Macfarlane GJ, Talamini R, Barzan L, Canova C, Simonato L, Lowry R, Conway DI, McKinney PA, Znaor A, McCartan BE, Healy C, Nelis M, Metspalu A, Marron M, Hashibe M, Brennan PJ (2009) Alcohol-related cancers and genetic susceptibility in Europe: the ARCAGE project: study samples and data collection. Eur J Cancer Prev 18(1):76–84

Liu K, Stamler J, Dyer A, McKeever J, McKeever P (1978) Statistical methods to assess and minimize the role of intra-individual variability in obscuring the relationship between dietary lipids and serum cholesterol. J Chronic Dis 31:319–418

Long JS (2009) The workflow of data analysis using stata. Stata Press, College Station

Lumsden J (2005) Guidelines for the design of online-questionnaires. National Research Council of Canada, NRC/ERB-1127, NRC 48231

Lynge E, Afonso N, Kaerlev L, Olsen J, Sabroe S, Ahrens W, Eriksson M, Guénel P, Merletti F, Stengrevics A, Suarez-Varela M, Costa-Pererra A, Vyberg M (2005) European multi-centre case-control study on risk factors for rare cancers of unknown aetiology. Eur J Cancer 41(4):601–612

Miettinen OS (1985) Theoretical epidemiology: principles of occurrence research in medicine. Wiley, New York

Moulin JJ, Wild P, Romazzini S, Lasfargues G, Perdrix A, Peltier A, Bozec C, Duguerry P, Pellet F (1998a) Lung cancer risk in hard metal workers. Am J Epidemiol 148:241–248

Moulin JJ, Clavel T, Chouaniere D, Massin N, Wild P (1998b) Implementation of ISO 9002 for research in occupational epidemiology. Accredit Qual Assur 3:488–496

National Cohort (2012) http://www.nationale-kohorte.de/index_en.html. Accessed 3 Sept 2012

Nurses' Health Study I, II, III (2011) http://www.channing.harvard.edu/nhs/. Accessed 3 Sept 2012

Olsen J, Christensen K, Murray J, Ekbom A (2010) An introduction to epidemiology for health professionals. Springer Series on Epidemiology and Public Health, vol 1. Springer, Heidelberg/New York

Olsson A, Kromhout H, Agostini M, Hansen J, Lassen CF, Johansen C, Kjaerheim K, Langård S, Stücker I, Ahrens W, Behrens T, Lindbohm ML, Heikkilä P, Heederik D, Portengen L, Shaham J, Ferro G, de Vocht F, Burstyn I, Boffetta P (2010) A case-control study of lung cancer nested in a cohort of European asphalt workers. Environ Health Perspect 118(10):1418–1424

Pearl J (1995) Causal diagrams for empirical research. Biometrika 82:669–710

Pearl J (2000) Causality: models, reasoning, and inference. Cambridge University Press, Cambridge

Peplies J, Günther K, Bammann K, Fraterman A, Russo P, Veidebaum T, Tornaritis M, Vanaelst B, Mårild S, Molnár D, Moreno LA, Ahrens W, IDEFICS Consortium (2011) Influence of sample collection and preanalytical sample processing on the analyses of biological markers in the European multicentre study IDEFICS. Int J Obes 35(Suppl 1):S104–S112

Peters S, Kromhout H, Portengen L, Olsson A, Kendzia B, Vincent R, Savary B, Lavoué J, Cavallo D, Cattaneo A, Mirabelli D, Plato N, Fevotte J, Pesch B, Brüning T, Straif K, Vermeulen R (2013) Sensitivity analyses of exposure estimates from a quantitative job-exposure matrix (SYN-JEM) for use in community-based studies. Ann Occup Hyg 57:98–106

Pohlabeln H, Wild P, Schill W, Ahrens W, Jahn I, Bolm-Audorff U, Jöckel K-H (2002) Asbestos fiber years and lung cancer: a two-phase case-control study with expert exposure assessment. Occup Environ Med 59:410–414

Reynolds RA, Woods R, Baker JD (eds) (2007) Handbook of research on electronic surveys and measurements. IGI Global, Hershey. doi:10.4018/978-1-59140-792-8

Rothman KJ, Greenland S (1998) Modern epidemiology, 2nd edn. Lippincott-Raven, Philadelphia

Rothman KJ, Greenland S, Lash TL (2008) Modern epidemiology, 3rd edn. Lippincott Williams & Wilkins, Philadelphia

Rouch I, Wild P, Fontana JM, Chouaniere D (2003) Evaluation of the French version of EUROQUEST: a questionnaire for neurotoxic symptoms. Neurotoxicology 24:541–546

Rylander R (2002) Endotoxin in the environment – exposure and effects. J Endotoxin Res 8: 241–252

Sandman PM (1991) Session III. Ethical considerations and responsibilities when communicating health risk information. Emerging communication responsibilities of epidemiologists. J Clin Epidemiol 44:S41–S50

Sauleau EA, Wild P, Hours M, Leplay A, Bergeret AR (2003) Comparison of measurement strategies for prospective occupational epidemiology. Ann Occup Hyg 47:101–110

Savitz DA (2003) Interpreting epidemiologic evidence. Strategies for study design and analysis. Oxford University Press, New York

Schwalbe K (2012) An introduction to project management, 4th edn. Kathy Schwalbe LLC, Minneapolis

Stansfeld S, Crombie R (2011) Cardiovascular effects of environmental noise: research in the United Kingdom. Noise Health 13(52):229–233

UK Biobank (2007) Protocol for a large-scale prospective epidemiological resource. Protocol No: UKBB-PROT-09–06 (main phase, amendment one final, 21 March 2007) UK Biobank Coordinating Centre, 1 & 2 Spectrum Way, Adswood, Stockport, Cheshire SK3 0SA

Wacholder S, McLaughlin JK, Silverman DT, Mandel JS (1992a) Selection of controls in case-control studies I: principles. Am J Epidemiol 135:1019–1028

Wacholder S, Silverman DT, McLaughlin JK, Mandel JS (1992b) Selection of controls in case-control studies II: types of controls. Am J Epidemiol 135:1029–1041

Wacholder S, Silverman DT, McLaughlin JK, Mandel JS (1992c) Selection of controls in case-control studies III: design options. Am J Epidemiol 135:1042–1050

Waksberg J (1978) Sampling methods for random digit dialing. J Am Stat Assoc 73:40–46

White E, Armstrong BK, Saracci R (2008) Principles of exposure measurement in epidemiology: collecting, evaluating and improving measures of disease risk factors, 2nd edn. Oxford University Press, Oxford/New York

WHO (2008) Handbook: good laboratory practice (GLP): quality practices for regulated non-clinical research and development, 2nd edn. World Health Organization, Geneva; also available at http://www.who.int/tdr/publications/documents/glp-handbook.pdf. Accessed 21 Sept 2012

Wild P, Moulin JJ, Ley FX, Schaffer P (1995a) Mortality from cardiovascular diseases among potash miners exposed to heat. Epidemiology 6:243–247

Wild P, Refregier M, Auburtin G, Carton B, Moulin JJ (1995b) Survey of the respiratory health of the workers of a talc producing factory. Occup Environ Med 52:470–477

Wild P, Leodolter K, Réfrégier M, Schmidt H, Bourgkard E (2008a) Effects of talc dust on respiratory health: results of a longitudinal survey of 378 French and Austrian talc workers. Occup Environ Med 65:261–267

Wild P, Andrieu N, Goldstein AM, Schill W (2008b) Flexible two-phase studies: a new variant to increase study efficiency to detect effects of rare exposures. Epidemiol. Perspect Innov 5(4): 1–11

Wolff MS, Toniolo PG, Lee EW, Rivera M, Dubin N (1993) Blood levels of organochlorine residues and risk of breast cancer. J Natl Cancer Inst 85:648–652

World Medical Organization (1996) Declaration of Helsinki. Br Med J 313:1448–1449

Quality Control and Good Epidemiological Practice 14

Gila Neta, Jonathan M. Samet, and Preetha Rajaraman

Contents

G. Neta (✉)
Division of Cancer Control and Population Sciences and Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA

J.M. Samet
Director, University of Southern California (USC) Institute for Global Health, Professor and Flora L. Thornton Chair, Department of Preventive Medicine Keck School of Medicine of USC

P. Rajaraman
South Asia Region Center for Global Health, US National Cancer Institute, New Delhi, India

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_14,

14.1 Introduction

The use of data is fundamental in epidemiology. Epidemiological research on causation uses data in a search for the true nature of the relationship between exposure and disease. Increasingly, research with biomarkers of susceptibility, exposure, and response seeks to characterize mechanisms of disease causation. Research on the consequences of interventions seeks an unbiased assessment of the effects of independently varying factors on the outcome measure(s).

One of the most rewarding moments for a researcher is obtaining the preliminary results from his or her study. However, the question "do I believe what I see?" should immediately come to mind. The answer to this question is determined in large part from the more mundane but critical question of how good is the quality of the data, rather than from the elegance of the scientific design and methods. Errors that occur during study population selection or in the measurement of study exposures, outcomes, or covariates can lead to a biased estimate of the effect of exposure on risk for the disease of interest. Misclassification of exposure or disease that occurs randomly between all study participants decreases the power of the study to detect an association where it exists. Data collection that is differentially biased may have more severe consequences and can lead to an incorrect assessment of the relationship between exposure and disease.

The inherently important issue of study quality is becoming of even greater consequence as the findings of epidemiological studies gain in impact and the field of epidemiology gains wider acceptance as an essential element of biomedical research (Hiatt 2010; Samet 2000; Samet and Lee 2001). Results of epidemiological studies are routinely reported in the media, receiving widespread attention because the findings have evident relevance to the populace. Epidemiological evidence is also used to inform regulatory and legislative policy making (Goldman 2001). The decision to set airborne standards for particulate matter in the United States, for example, was largely fueled by evidence from epidemiological studies (Bachmann 2007; Greenbaum et al. 2001). Epidemiology often figures prominently in litigation, where the study methodology can become a point of debate (Bryant and Reinert 2001; Goldman 2001; Michaels 2008). Given the significance of epidemiological evidence for decision making, the results of epidemiological studies often face close scrutiny and questions may be raised about every aspect ranging across data quality, study methods, study conduct, data analysis, and interpretation of findings.

Even if external questioning and auditing are not anticipated, the researcher nonetheless faces the responsibility of assuring the quality of the study and preventing the widespread dissemination of misleading or incorrect information. For example, findings from several cohort studies on air pollution and mortality figured prominently in a 1997 decision by the US Environmental Protection Agency (EPA) to promulgate a new standard for airborne particulate matter (Bachmann 2007). The great weight given to the data by the EPA led to a call for access so that others could check and analyze the data. An extensive reanalysis of the data was carried out, including validation of elements of the original data as well as replication and extension of the original analyses by an independent group (Krewski et al. 2000; Samet et al. 1997). The controversy surrounding the use of data from the air pollution cohort studies eventually led to a Congressionally mandated requirement for sharing data with policy implications that have been collected with support by federal funds (Data Quality Act 2000, Public Law 106–554, Sect. 515a).

Many hypotheses of current interest in epidemiological studies call for the incorporation of data from multiple centers and involve collection of data from large populations according to centrally standardized protocols. Data sharing has also become more common, and approaches to doing so for larger grants in the United States have been mandated by the National Institutes of Health (Samet 2009). In order to enhance statistical power, data from individual studies are often pooled, or summary results are combined using meta-analysis. The establishment of networks or cohort consortia also allows for enhanced statistical power to address increasingly complex research questions, particularly in the general area of molecular and genetic epidemiological studies. These approaches to data utilization place a further demand for meticulous study documentation so that data from a study are readily usable by persons other than the original investigators.

General methods have long been available for assuring the quality of data. The idea of creating a high quality end product using process improvement initially emerged in the context of industrial business models. Early efforts at delivering quality products to customers were based on inspecting products at the end of a factory line and eliminating those products that did not meet standards ("quality control"). The idea of improving all procedures that affect the quality of the manufactured products ("quality assurance") represented a fundamental shift in paradigm for industrial manufacture. Incorporating quality considerations into the process rather than the product has since gained widespread acceptance in the business and engineering communities (International Organization for Standardization 2003).

Although there is a vast literature on quality control in general, the issue has not received much formal attention in the epidemiological setting. Within epidemiology, much of the writing on data quality and good epidemiological practice is focused on the conduct of clinical trials and pharmacoepidemiology (Canner et al. 1983; Cooper 1986; Dischinger and DuChene 1986; DuChene et al. 1986; Etminan et al. 2006; Gassman et al. 1995; Hilner et al. 1992; Knatterud et al. 1998; Meinert and Tonascia 1986; Neaton et al. 1990; Vantongelen et al. 1989). While clinical and laboratory guidelines can easily be modified to make them more applicable to observational studies, few sources specifically address quality issues for the most common epidemiological study designs. In an early attempt to bridge this gap, the Epidemiology Task Group of the Chemical Manufacturers Association (CMA) compiled a set of guidelines for good epidemiology practice for occupational and environmental epidemiological research (Cook 1991; The Chemical Manufacturers Association's Epidemiology Task Force 1991). An overview of data quality issues for epidemiological studies is also provided by Szklo and Nieto (2000) and by Whitney and colleagues (1998). Methods to improve data quality in medical registries are reviewed by Arts et al. (2002).

This chapter provides a general overview of data quality and guidelines for good practice in epidemiological research. The fundamental premise is that quality considerations should be integrated into every phase of the study from initial hypothesis formulation to the final publication of findings and archiving of data. Obtaining data completely free from error clearly would be prohibitively expensive and often impossible. The goal is therefore not error-free data, but rather planning and implementing cost-effective procedures that guarantee the validity of the primary results to an acceptable degree. The epidemiological researcher needs to be able to gauge the extent of any errors and assess the consequences for interpretation of data analyses. The idea of "quality control" versus "quality assurance" is carried over from the industrial management literature into the epidemiological literature, with a distinction made between activities that take place prior to data collection (quality assurance) and activities that occur during and after data collection to correct data errors (quality control).

The ubiquitous nature of quality issues, both in terms of where these issues can arise and how they affect study results, can be captured by an extended metaphor. In an article describing the causation of bias, Maclure and Schneeweiss present the idea of an "Episcope" through which an epidemiologist views a putative association between a causal agent and morbidity. Just as a user of a large telescope would be skeptical about whether and how image degradation exists, an epidemiologist should think about how and why an observed association between exposure and disease might be biased (Maclure and Schneeweiss 2001). A similar idea can be applied to data quality. As published study results are viewed through a "datascope," a discerning epidemiologist should be wary of how the final image (the published results) may have been distorted by quality considerations during the design, conduct, and dissemination of the study. Working backward, the observer might ask a string of questions, such as "Were the observed results more likely to be published because they were positive findings? Based on the analysis, were published inferences appropriate? Were the methods of analysis suitable? Were data keyed in correctly? Has the data been collected appropriately? Was an appropriate population defined?" Each of these questions points to one or more study quality issues. Using the metaphor of the datascope, we will highlight the main issues regarding study design and conduct, and present ways in which to improve epidemiological practice and data quality.

14.2 The Datascope

Imagine, for a moment, that published study results can be viewed only through a large telescope. As you peer into the lens, the initial picture is barely discernible. On your right is a panel with focusing controls. The first dial allows you to adjust out any distortion caused by publication bias. When you optimize this dial, the image becomes slightly clearer. The next control allows you to tune out faulty inferences. Again, you turn this knob to make the image somewhat clearer.

The process continues until the results are finally sharply focused. Although we do not literally look through a telescope every time we view the results of a study, we are in fact looking at an association that may well be "out of focus," depending on how well the study was designed, conducted, and interpreted. Errors in the measurement of the exposure, outcome, or other covariates can be thought of as unfocused datascope controls that contribute to degradation of the final image.

Let us consider, in some more detail, the datascope controls that manipulate sources of measurement error. The farthest dials from the observer are located in the planning phase of a study and influence purely "quality assurance" activities. For a more in-depth discussion of the planning stage, see chapter ▶Design and Planning of Epidemiological Studies of this handbook. Errors occurring at the study planning stage are summarized below.

- *Errors in Study Conception*
  If the study rationale and design are not carefully formulated, the rest of the study could be rendered completely irrelevant. Errors in study conception include inadequate literature review, consideration of an inappropriate study design, and failure to plan the validation of exposure or outcome variables.
- *Errors in the Selection, Design, or Procedures for Use of Instrument Measuring Exposure*
  The instrument selected for study exposure measurement might not cover all sources of the active agent. Conversely, the measurement instrument might include sources of exposures that are not biologically relevant or measure exposure for a time period that is etiologically unimportant. In survey instruments, the phrasing of questions or instructions could lead to misunderstanding or bias (cf. chapter ▶Epidemiological Field Work in Population-Based Studies of this handbook for a discussion of bias). Insufficient detail in the protocol for instrument use or inadequate consideration of a standardized method for dealing with unusual situations can lead to collection of poor quality data.
- *Inadequate Training of Study Personnel*
  Even if study procedures are very well defined, inadequate and non-standardized training of data collection staff in the application of these procedures can introduce errors in the data (cf. chapter ▶Epidemiological Field Work in Population-Based Studies of this handbook)

The next set of controls is activated during the conduct of a study and includes activities that generally fall under the categories of quality assurance as well as quality control. For instance, validation studies of instruments and equipment ensure that collected data will be accurate (quality assurance) but can also be used to correct errors in data (quality control). Sources of exposure measurement error that can occur during data collection are described in detail elsewhere (White et al. 2008) and summarized below.

- *Improper Execution of the Study Protocol*
  Errors related to study protocol execution include the misinterpretation of, or deviation from, standard operating procedures by study technicians. Mistakes in interpretation of the study protocol often arise from poor clarity of the manual of

operations or inadequate training of study personnel. For example, if the standard operating procedure states that a fasting blood glucose level should be measured but does not specify the time required to have elapsed after the last meal, the interpretation of "fasting" may differ from technician to technician. Errors in data can also result from improper handling of biological specimens or the failure of subjects to read or understand instructions in self-administered questionnaires. For some biological assays, meticulous handling of a blood specimen may be needed.

- *Errors Related to Study Participants and Intra-individual Variability*
  Subjects may have poor recall of past exposures or allow recent exposure to influence their memory of past exposure. Individuals also tend to overreport socially desirable behaviors, such as exercise, and underreport socially undesirable habits such as smoking. Additionally, short-term variability in the biological characteristics of a subject can lead to unrepresentative measurements of exposure or outcome. For example, differences in the levels within or across individuals of an exposure biomarker measured at a specified time after exposure are likely to be due partially to individual differences in metabolizing the agent of interest.
- *Changes in the Accuracy of Measurements over Time*
  Failing to standardize and recalibrate laboratory equipment is likely to introduce data drift as calendar time progresses. In long-term studies, the instrument used for measurement may change over time, and the agent of interest in biological specimens may be subject to degradation. Also, as the study personnel gets more experienced through the course of the study, changes in the handling of procedures and instruments may occur.
- *Mistakes in Data Processing*
  Data that are recorded inaccurately, illegibly, or incompletely are very difficult to correct after the fact. Transcription of the data to electronic files introduces more chances for error, both within a study site and between field sites and the data coordinating center. At the coordinating center, programming or procedural errors may corrupt the database or modify data inappropriately. Errors can also be introduced by undocumented changes or modifications to a local or central database.

The final panel of controls on the datascope, closest to the observer, consists of purely "quality control" dials, which influence study quality after the data have been collected. Examples of these errors are presented below.

- *Inappropriate Data Analysis*
  If data analysis is not preceded by familiarization with the nature of the data, the chosen analyses may not be appropriate. Specifying the wrong model for analysis, for instance, can lead to completely erroneous results and inference.
- *Poor Reporting of Data*
  Omitting the results of important data analyses or presenting unnecessary information can obfuscate the study results. Lengthy, verbose explanations and poorly labeled graphs and figures add to the confusion. Inappropriate inference given the study results can also be misleading.

In order to achieve the highest quality data possible, each of the sources of error described in the planning, design, conduct, and conclusion of a study should be minimized. Conceptually, this can be thought of as turning the appropriate datascope dial to obtain the best image possible.

A review of the different sources of error that can occur during study planning, designing, and conducting informs the datascope user as to where he or she can affect final data quality. The ultimate goal is to optimize the datascope dials in order to minimize error and achieve the clearest possible picture of the study results. In the rest of this chapter, we present aspects of quality control and good epidemiological practice that can reduce data error. The chapter will follow the same organization as the datascope control panels, beginning with the planning phase of a study, moving onto quality considerations during study conduct, and finally describing activities that occur after data collection. Where applicable, the working of the datascope will be illustrated using the example of measuring blood pressure in a hypothetical study whose main research question is whether elevated blood pressure leads to increased risk of coronary disease.

14.3 Quality Considerations in the Planning Phase

14.3.1 Protocol

The development of a comprehensive study protocol is essential to good epidemiological practice. The *study protocol* is a narrative document that describes the general design and procedures used in the study. It can be distinguished from the study *manual of operations* (Sect. 14.3.3) by its generality and absence of specific details for day-to-day study conduct. The study protocol assists the staff in understanding the context in which their specific activities occur. A well-designed study protocol can, and should, guide all aspects of the study. In general, a protocol would include the following sections: a short descriptive title, a description of performance sites and personnel, a description of background and significance, results of preliminary studies, study design and methods, a time line for completion of major tasks, ethical considerations, and references. Quality assurance and quality control should be addressed in each relevant section of the protocol and also summarized in a separate section. Although restrictions or recommendations provided in the guidelines for research grants applications for the US National Institutes of Health may not be applicable to grants funded through other mechanisms, these guidelines nevertheless provide useful suggestions for creating study protocols (US Department of Health and Human Services 2009). Recommendations for protocol write-up are also included in the Guidelines for Good Epidemiology Practices for Occupational and Environmental Epidemiological Research (The Chemical Manufacturers Association's Epidemiology Task Force 1991). The typical sections of a study protocol are summarized in Table 14.1 (see also chapter ▶Design and Planning of Epidemiological Studies of this handbook).

Table 14.1 Guidelines for preparation of a study protocol (Adapted from the guidelines for good epidemiology practices, epidemiology task group (The Chemical Manufacturers Association's Epidemiology Task Force 1991))

Section	Guidelines for good epidemiological practice
Title	Descriptive and to the point
Names, titles, degrees, addresses and affiliations of the study director, principal investigator, and all coinvestigators	Possible conflict of interest should be identified and resolved
Name(s) and address(es) of the sponsor(s)	Possible conflict of interest should be identified and resolved
Proposal abstract	Informative and succinct
Proposed study tasks and milestones	Timetable should be realistic and identify possible sources of delay
Statement of research objectives, rationale, and specific aims	Clearly state the purpose of the investigation, describe whether the study will be hypothesis generating or hypothesis testing, and whether the study will confirm previous findings or result in new findings
Critical review of the relevant literature	Include animal, clinical, and epidemiological studies Do not restrict search to electronic databases (e.g., PubMed, TOXLINE), older articles might be missed Describe the occurrence of exposure and outcome variables Identify potential confounders and effect modifiers Identify gaps in current knowledge
Description of the research methods	Describe the overall research design and why it was chosen. Consider alternative designs Define exposure and outcome variables, and identify data sources for these and other variables of interest. Check whether the measure of exposure represents the biologically active agent and etiologically important time period Calculate the projected study size and statistical power (if appropriate) Describe procedures for collecting data Provide a detailed description of the methods of analysis Define how exposure and outcome variables will be categorized for analysis State how confounders and effect modifiers will be treated in the analysis Outline the major strengths and limitations of the study design Provide criteria for interpreting the study results, including ways of assessing statistical, clinical, and biological significance
Description of plans for protecting human subjects	Describe risks and benefits of participating in the study If appropriate, provide plans for obtaining informed consent Describe procedures for maintaining confidentiality of subjects and data

(*continued*)

Table 14.1 (Continued)

Section	Guidelines for good epidemiological practice
Description of quality assurance and control	Describe for all phases of the study
Resources required to conduct the study	Detail the expected time, personnel, and equipment required for the study
Bibliographic references	Include all relevant references
Addenda, as appropriate	Examples of useful addenda include copies of collaborative agreements, institutional approvals, informed consent forms, and questionnaires
Dated protocol review and approval sign-off sheet	Document dated amendments to the protocol

Improving the datascope image by choosing appropriate measures of hypertension

In the planning phase of the study, investigators should make provision for collection of appropriate measures of hypertension. While clinicians favor the diagnosis and treatment of hypertension in terms of diastolic blood pressure elevation, data from the Framingham Study in Massachusetts indicate that systolic blood pressure is a better predictor of disease outcome (Kannel 2000). Additionally, ambulatory blood pressure can be measured with an automated device so that multiple measurements can be made across the course of typical activities. Studies show that such recordings provide information predictive of disease risk beyond that obtained with measurements made at a single assessment (Clement et al. 2003).

14.3.2 Documentation of Operations and Procedures

The consistency and validity of study data are greatly enhanced by the establishment and application of *standard operating procedures* for routine data collection tasks (a *standard operating procedure* is defined here as a standardized method or process for conducting a routine research procedure). If standard procedures have been well described, variability is likely to be much lower across study sites, interviewers, or technicians. Uncorrected variability introduced by interviewers or technicians can decrease study power (Blomgren et al. 2006; Edwards et al. 1994, 1998).

Standard procedures should be clearly described for all study procedures, including (but not limited to) raw data collection, coding of death certificates, assessment of error rates, and management of archived data. Each standard operating procedure should state the purpose of the procedure, provide a detailed description of the procedure including forms and equipment to be used, and either designate the person responsible for the procedure, or explain what training will be needed

(The Chemical Manufacturers Association's Epidemiology Task Force 1991). The US Toxic Substances Control Act (TSCA) standard for Good Laboratory Practices provides detailed quality control and quality assurance guidelines for the collection of laboratory samples (US Environmental Protection Agency (EPA) 1989). Once the various study operating procedures are established, they should be integrated and summarized in the form of a study *manual of operations*. The *manual of operations* is a document or collection of documents that completely describes the procedures used in a study center. Developing a *study handbook*, which contains a series of tables, charts, figures, and specification pages that outline the main design and operating features of a study (largely without the use of a written narrative), is a useful first step in the development of the manual of operations and can also act as a quick reference for study personnel. The study protocol, handbook, and manual of operations should be reviewed for clarity and completeness.

Since the initial version of the manual of operations is almost certain to contain some errors, pretesting of the manual prior to finalization is essential. All aspects of the study protocol should be tested on a population similar to the one that will be studied, including the administration of surveys, sending of samples to laboratories, and the generation of and response to quality control reports. Refinements to the protocol that are identified from the pilot study can be incorporated into the final study manual of operations.

Improving the datascope image using standard operating procedures

Inter- and intra-technician variation in blood pressure readings viewed under the datascope can be reduced by clear and detailed descriptions of the method of measurement. Application of a standard operating procedure can also reduce variability in blood pressure measurement within a subject. Specifying details such as how the study participant should be seated, which arm the cuff should be applied to, and how long the study participant should remain quiet before the reading is taken can reduce the influence on the study measurement of factors that affect an individual's blood pressure.

14.3.3 Personnel: Training and Certification

Integral to study conduct is the availability of personnel with the necessary education, experience, and training to perform assigned functions. The planning stage of the study is the appropriate time to consider what personnel will be required, what kind of training will be necessary, and how often training should occur. Job descriptions should be written for each individual who will be supervising or engaging in study conduct. For jobs that require training, procedures for initial and retraining of personnel should be established. Retraining may be necessary if substantial time has elapsed since the initial training, if a technician is found to be introducing a systematic error into the data, or if the study protocol changes. For each of the study personnel, a summary of relevant training and experience, including study certification and recertification, should be maintained and kept up to date.

Consistency in the training of personnel across sites improves comparability of data collection across different study sites. This training can be centralized or site specific. Often, a combination of both approaches is used (see Sect. 14.4.1 for more detail). Study personnel should be required to follow standard operating procedures. If training will be difficult or time-consuming, it is prudent to train at least two individuals for each task in case one of the trained technicians leaves the study. Certification standards should be set and might include completion of a specified number of tests for key procedures, including some under observation.

Aside from the obvious benefit of consistency in data collection, training study personnel also increases the interviewers' or technicians' perceived value of the data that are being collected. This may influence the amount of care taken in following the protocol. Some studies use computer instruction, DVDs, or teleconferencing to reduce the costs associated with training. While the use of computer or DVD training is convenient, these methods lack some of the benefits of face-to-face training, such as the opportunity for staff members to share ideas and the opportunity for scientific presentations that remind personnel of the importance of their work (Whitney et al. 1998).

Improving the datascope image by training and certification

Some of the variation in blood pressure measurements viewed under the datascope could be due to a technician-measured blood pressure in a different way each time he or she takes a measurement or if different technicians have different ways of reading the same measurement. One way to minimize these sources of error and improve the datascope image is to train and certify study technicians. In the MRFIT (Multiple Risk Factor Intervention Trial), technicians were trained in taking blood pressure measurements using training tapes and a double stethoscope (Dischinger and DuChene 1986). The training tapes consisted of two recordings of the Korotkoff sounds for 12 subjects. The first tape of Korotkoff sounds was used for training and the second for testing. A video-training film that presented 12 blood pressure readings, with sufficient time to determine and record systolic and diastolic blood pressure after each reading, was also used. Finally, supervisors and trainees took simultaneous measurements of three subjects using a double stethoscope. The differences in the readings of the trainer and technician had to fall below a certain criterion for the trainee to pass. Technicians were certified after completing the training tapes, passing a written test on procedures for taking blood pressure measurement, and passing the double stethoscope test. Recertification was required at regular intervals or if examination of collected data indicated that a technician had a bias with respect to other technicians in a clinic.

14.3.4 Data Collection Forms and Instruments

Exposure and outcome measures for epidemiological studies can be collected in a variety of ways. Methods of data collection include mailed self-administered questionnaires, interviewer-administered questionnaires, measures of blood or other tissues, physical measures, medical tests, use of medical or exposure records, or sampling for environmental contaminants (White et al. 1998, 2008). Most studies use more than one method of data collection.

The use of data that have been collected already ("secondary data") has the key advantage that the data already exist. Studies using secondary data are thus likely to be more cost and time-efficient than studies with primary data collection. Sources of secondary data, such as population-based registries, often allow for a much larger sample size and are more representative of the general population (Hearst and Hulley 1988). A substantial disadvantage of using existing data, however, is that the collected data may not adequately address the particular research question of interest. An additional drawback is that the method of collection and the quality of the secondary data are not under the researcher's control. For this reason, researchers using secondary data should carefully review data documentation and evaluate the quality and validity of these data to the extent possible (Bray and Parkin 2009; Clive et al. 1995; Gissler et al. 1995; Goldberg et al. 1980; Horbar and Leahy 1995; Maudsley and Williams 1999; Sorensen et al. 1996; Wyatt 1995). For details on the use of secondary data, see chapter ▶Use of Health Registers of this handbook.

Most epidemiological studies collect some or all of their data using phone, mail, or self/interviewer-administered questionnaires. Data from such questionnaires, however, can be subject to various sources of bias. For instance, study participants filling out a self-administered questionnaire might report socially acceptable rather than strictly accurate results. Moreover, ways of responding to the survey may differ between participants in the study, depending on factors such as the age, gender, or racial/ethnic group of the participant. Conversely, participants may respond differently to interviewers of different age, gender or ethnic background. For further details, see chapter ▶Epidemiological Field Work in Population-Based Studies of this handbook. Multicenter studies encounter the additional problem of differences in data collection between study centers. In long-term studies, these biases can change over time. Smoking, for example, is generally less socially acceptable today than it was 20 years ago in the United States and consequently more likely to be underreported (Ling and Glantz 2002; O'Connor et al. 2006; US Department of

Health and Human Services 2004). The acceptability of smoking also varies by ethnic groups, which may in part explain the fact that African-American high school seniors are far less likely to smoke than are white seniors (Wallace et al. 2002). Measurement error that occurs because of the use of a survey instrument can be minimized by careful design and pretesting of the survey and the application of standardized interviewing techniques.

The main objective of survey design is to allow the efficient collection of data that are valid, reliable, and complete. Standardizing forms within a study is important for internal validity. Consistency of forms across studies allows more meaningful comparison with other studies and also makes the study results more generalizable. Both internal and external form standardization can be achieved by the use of preexisting validated study instruments. Examples of validated questionnaire instruments include the American Thoracic Society questionnaire to assess respiratory symptoms (Comstock et al. 1979), the Willett food frequency questionnaire (Willett et al. 1985), the Piper Fatigue Scale (Piper et al. 1989, 1998) to assess fatigue in cancer patients, the Wechsler Preschool and Primary Scale of Intelligence (Wechsler 1989), and the International Physical Activity Questionnaire (Craig et al. 2003). If a validated instrument is not readily available, several sources in the literature provide guidelines for questionnaire design to maximize clarity and ease of administration. These include recommendations for physical format, as well as instructions on how to word the text of instructions and questions (Dillman 1978; Holford and Stack 1995; Knatterud et al. 1998; Meinert and Tonascia 1986; Wright and Haybittle 1979a, b, c). Studies that enroll participants of different ethnic groups may need to accommodate different languages by using interpreters or by having translated versions of the questionnaire. However, a question might change subtly upon translation, and data generated from different languages may not be entirely comparable. For this reason, independent back-translation of questions to the original language is strongly recommended. An example of the need for back-translation is provided by data from a health survey which showed lower data reliability of data for Hispanics interviewed in Spanish than for Hispanics interviewed in English when no back-translation was done. An independent back-translation aimed at creating a linguistically equivalent version to the Spanish version indicated several instances in which the two versions were idiomatically different and appeared to have affected the seriousness with which the interview situation was perceived, in turn leading to response discrepancies (Berkanovic 1980).

Pretesting of the survey instrument on a population similar to the study population allows the detection of flaws in the survey design and instrument before full-scale data collection begins. Separate analysis of pretest data by language version, for example, might identify problems in translation. In the Hypertension Prevention Trial, which was designed to test the effectiveness of changes in dietary intake of calories, sodium, and potassium, a test cohort of 78 participants was enrolled and used for the testing of forms and procedures. Data that were generated from the test cohort were used to identify problems in survey design and collection and were not analyzed with results from the main study (Prud'homme et al. 1989).

Accuracy and consistency are also important for laboratory or clinical equipment. The study should be planned so that all study personnel and sites begin by using identical equipment. In anticipation of measurement drift over time, procedures to maintain and recalibrate equipment should be established. In the Sleep Heart Health Study (Quan et al. 1997), overnight sleep data were collected from subjects using a portable monitor. Sites were notified to have the monitor evaluated and procedures assessed when less than 85% of results scored by the monitor were of "good" or better quality (Whitney et al. 1998). Standard and random zero sphygmomanometers for blood pressure measurement in the MRFIT study (Kjelsberg et al. 1997) were maintained and calibrated according to a regular schedule and subject to standard checks at least every other month in the case of the standard sphygmomanometer and every week in the case of the random zero instrument (Dischinger and DuChene 1986).

As more advanced technology becomes available to measure an exposure or outcome, there may be justification to update study equipment. For example, mobile phones and other personal digital assistants (PDAs) are increasingly being used for data collection (Aanensen et al. 2009; Kwok 2009). In such cases, data should initially be collected using both the old and new equipment to establish the comparability of the two instruments since a change may introduce subtle differences that are only apparent as substantial data are collected using the new approach.

14.3.5 Planning to Achieve a High Response Rate

In order to curtail the possibility of bias and increase the generalizability of study results, it is important to achieve the highest response rate possible (Gordis 2000; Wacholder et al. 1992). A systematic review of 372 trials found that factors which more than doubled the odds of response to surveys were the inclusion of a monetary incentive with the questionnaire, the designing of surveys to be of more interest to participants and the use of registered mail (Edwards et al. 2007). Other factors which have been reported to increase response rate are shorter questionnaire length (Dillman 1978; Eaker et al. 1998; Hoffman et al. 1998; Kalantar and Talley 1999; Kellerman and Herold 2001; Little and Davis 1984; Martinson et al. 2000; Spry et al. 1989), personalizing questionnaires (Maheux et al. 1989), using colored ink (Edwards et al. 2007), contacting participants before sending questionnaires, providing stamped return envelopes (Choi et al. 1990), and using written or telephone reminders (Asch et al. 1997). Questionnaires originating from universities are more likely to be returned than questionnaires from other sources, whereas surveys eliciting information of a sensitive nature are less likely to be returned.

While the use of a monetary incentive is probably the factor that has been shown most consistently to increase response rates (Gibson et al. 1999; Gilbart and Kreiger 1998; Hoffman et al. 1998; Kellerman and Herold 2001; Martinson et al. 2000; Parkes et al. 2000; Perneger et al. 1993), increasing the amount of the incentive results in diminishing returns of questionnaires after a certain point (Halpern et al. 2002; James and Bolstein 1992; Spry et al. 1989). In the

United States, the $2.00 bill has been a cost-effective monetary incentive in the past (Asch et al. 1998; Doody et al. 2003; Shaw et al. 2001). Making a prepayment of the incentive appears to be more cost-effective than promising payment on completion of the questionnaire (Schweitzer and Asch 1995). Including the monetary incentive in the first mailing rather than in subsequent mailings has resulted in higher response rates (John and Savitz 1994). Non-monetary incentives, while reported to increase response rates over having no incentive, do not appear as effective as monetary incentives (Kellerman and Herold 2001; Martinson et al. 2000).

Contact rates generally tend to be lower for individuals who are young, male, black, of lower socioeconomic status, or employed full-time (Collins et al. 2000; Cottler et al. 1987; Moorman et al. 1999). In the context of a case-control study, response rates are often lower for controls (Moorman et al. 1999). Even within control groups, different types of controls have different response rates. For example, in the United States, controls chosen from Health Management Organizations have been shown to have a higher response rate than controls drawn from lists of licensed drivers (Slattery et al. 1995).

Long-term cohort studies, in addition to having to address response rates to study questionnaires, also face the issue of loss to follow-up. The loss of cohort members to follow-up is conceptually similar to response rate, in that loss to follow-up can constitute an important source of selection bias and also limit external validity. Participants may be lost to follow-up either because they drop out of the study of their own volition or because the study investigators lose track of them. As with other types of epidemiological studies, loss to follow-up can lead to reduced study power and may result in biased estimates of risk. Strategies for minimizing loss to follow-up include pre-enrollment screening of participants for willingness to participate in a long-term study, collecting names of personal contacts and proxies for participants, maintaining regular contact with study participants, using incentives for remaining participants, and maintaining tracking systems to follow participants (Hunt and White 1998; White et al. 2008). One must keep in mind, however, that populations comprised of volunteers are usually different from the population as a whole. In general, measures of relative risk are less affected by the lack of external generalizability than measures of absolute or attributable risk.

14.3.6 Validity and Reliability

The absence of bias in data measurement is called *validity* or accuracy. The precision, or reproducibility, of collected data is known as *reliability*. See Fig. 14.1 for an illustration of the distinction between validity and reliability.

14.3.6.1 Validity Studies

The capacity of a measure to capture the true value of the exposure, outcome, confounder, or modifier of interest in the study population is known as its validity.

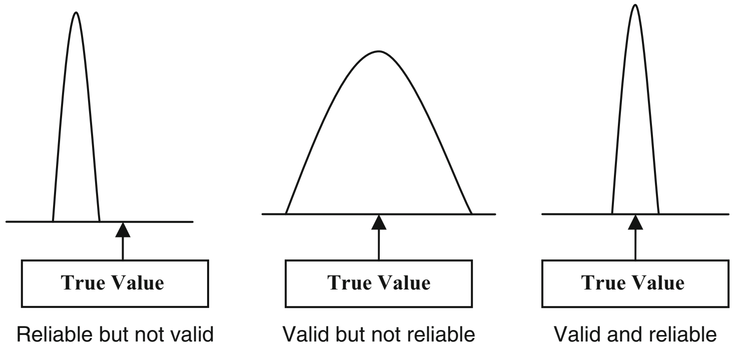


Fig. 14.1 Graphs of hypothetical test results illustrating the distinction between validity and reliability

While it is desirable to obtain the most accurate measurements possible of exposure or outcome, such measurements usually come at the price of increased cost, invasiveness, or time involvement. When faced with these constraints, epidemiologists often choose to collect less accurate measures of exposure.

The accuracy of the study's main method of exposure measurement can be assessed using validation studies which compare the study exposure measure to a more accurate measure of exposure ("gold standard"), either in a subsample of study participants or in a different population. For instance, evidence of validity can be provided by comparing study estimates of an environmental exposure to industrial hygiene measurements or biomarkers of exposure (Cherrie and Schneider 1998; Cherrie et al. 1987; Dosemeci et al. 1997; Hawkins and Evans 1989; Kipen et al. 1989; Kromhout et al. 1987; Tielemans et al. 1999). The comparison of reported nutrient intake on a questionnaire with a biochemical indicator provides another example of this approach (Ascherio et al. 1992; Johnstone et al. 1981; Post and Kromhout 1991; Sacks et al. 1986; Willett et al. 1983).

The establishment of a serum pool can facilitate validation of biological sample processing in the study. Study measurements can be compared with results from a "gold standard" external laboratory. If study measurements deviate randomly from the gold standard, the study result would be attenuated toward the null hypothesis. However, if deviations from the gold standard are found to vary according to the presence and level of important variables such as follow-up time or the exposure or outcome of interest, the study results may be biased.

Data from validation studies can, additionally, be used to account for uncertainty in the data analysis. Measurement error correction models can be developed that use validation study data to adjust the full data for measurement error (Holford and Stack 1995; Rosner et al. 1992; Spiegelman et al. 1997; Stram et al. 1999; Thompson 1990). In the Framingham Study (Dawber et al. 1951), for example, a small validation study was conducted to estimate the relationship between the surrogate measurement (food frequency) and the "true" measurement (diet record). Based on information from the main study relating the surrogate to disease outcome and information from the validation study relating true and surrogate exposure, corrected point estimates of risk were calculated (Rosner et al. 1992). For a general discussion of statistical methods to account for measurement errors see chapter ►Measurement Error of this handbook.

While validation studies may help to form a clearer picture of the true relationship between exposure and outcome, such studies are not without their own limitations. For one, the gold standard used for comparison may itself be subject to error, thus the term "alloyed" gold standard (Wacholder et al. 1993). While calibration methods for such alloyed gold standards have been described, these complex models cannot be applied in all situations (Kaaks et al. 2002; Spiegelman et al. 1997). A second limitation of validity studies is that participants in these studies are not always representative of all participants. Subjects who volunteer to take part in a validity study are likely to be more compliant than non-volunteers would have been. Additionally, feasibility constraints often limit validity studies to small sample sizes, which can lead to statistical imprecision.

14.3.6.2 Reliability Studies

Data variation can arise within study participants (*biological variability*) or due to variation in exposure assessment or physiological measurements introduced by study technicians. Blood pressure within an individual, for example, experiences short-term changes due to factors such as activity and mood. Different blood pressure measurements taken on the same individual are thus likely to vary for physiological reasons regardless of how accurately these measurements are made. Study technicians can add an extra component of variation to the measurements, either because a given technician reads a measurement in slightly different ways each time (*intra-observer variation*) or because different technicians read a measurement in different ways (*inter-observer variation*). Variability can also be introduced as samples degrade over time.

As illustrated in the paragraph above, variability in data can arise due to true change, measurement error, or random biological variation. The component of variability that the researcher is most interested in is the true change in study exposure that might influence the outcome under consideration. The separation of desired variability in the data (true change) from undesirable variability due to measurement error or random (biological) variation can be partially assessed by incorporating into the main study a series of substudies that are designed to assess the reliability of the study data.

Reliability studies can be used to assess various components of variability, such as the comparability of measurements taken by the same technician at a given visit, different technicians at a given visit, the same/different technician at different points of time, or the same/different technician using different instruments. Interobserver and intraobserver variability can be assessed using a set of calibration samples that are read several times by each technician and processed by multiple technicians. Biological variability can be assessed by having a single technician perform repeat studies on a subset of participants, although some amount of variability assessed in this way would be due to technician variability (Whitney et al. 1998). For durable data, such as X-ray films or dietary recall records, variation over time can be assessed by comparing evaluation of the same samples at different times in the study. In instances where samples are limited or perishable (e.g., blood or urine), a preselected set of "quality control" specimens should be set aside at the beginning of the study so that small amounts of these specimens can be periodically submitted for processing. Caution should be taken through careful planning at study onset to avoid specimen degradation from freeze thaw cycles. Technicians handling quality control samples should be unaware that the samples are different from other study samples being processed, in order to prevent differential handling.

Reliability studies which collect replicate measurements at the same point of time are useful in the identification of possible data errors, as well as in the calculation of more accurate measures of exposure. Averaging repeated measures has been recommended as an effective method of decreasing the measurement error associated with a single measurement (Canner et al. 1991; Holford and Stack 1995; White et al. 1998, 2008).

When validity is reported, the number of samples that are deemed unacceptable for analysis should be stated since this may indicate a bias in the remaining samples. The examination of whether reliability estimates differ according to relevant characteristics such as exposure, confounding factors, or outcome allows some assessment of whether differential misclassification is occurring in the data.

Improving the datascope image by obtaining a more valid exposure measure

Using the average of three blood pressure measurements taken on a study visit would result in a clearer picture of the individual's blood pressure than would a single blood pressure measurement.

14.3.6.3 Measures of Agreement

Quantifying the agreement between two different methods of measurement requires the use of some measure of agreement. The choice of statistic depends on the type of variables being compared and the purpose of the comparison (Table 14.2). Different measures of agreement and advantages and disadvantages of each measure have been reviewed by Szklo and Nieto (2000).

The basic measures of validity for binary categorical variables are sensitivity and specificity, for which the study value of the exposure or outcome is compared to the "true" value, measured by a more accurate method (Example 14.1). The *sensitivity* of a test is the ability to correctly identify those individuals who have the disease or exposure characteristic of interest. The test *specificity* is the ability to correctly

Table 14.2 Common measures of agreement, interpreted in the context of two separate technicians reading the same data

Statistic	Range	Type of data	Interpretation
Sensitivity and specificity	0–100%	Categorical	Sensitivity: ability to correctly identify individuals who have the disease or exposure characteristic of interest Specificity: the ability to correctly identify individuals who do not have the disease or exposure characteristic of interest
Overall percent agreement	0–100%	Paired categorical variables	The proportion of all readings that are categorized in the same way by two different observers Higher value means better agreement
Percent positive agreement	0–100%	Paired categorical variables	The proportion of all non-negative readings that are categorized in the same way by two different observers
Kappa statistic, κ	−1 to 1 (rarely below 0)	Paired categorical variables	The extent of agreement between two readers beyond that due to chance alone Higher value of kappa means better agreement
Weighted kappa	−1 to 1 (rarely below 0)	Paired categorical variables	The extent of agreement between two readers beyond that due to chance alone, allowing for consideration of partial agreement
Pearson's correlation, r	−1 to 1	Continuous ordinal variables	The degree to which a set of paired observations in a scatter diagram approaches the situation in which every point falls exactly on a straight line −1 is perfect negative correlation, 1 is perfect positive correlation
Spearman's correlation, r_s	−1 to 1	Non-parametric ordinal variables	The degree to which ranking of measurements is consistent between two readers
Intraclass correlation coefficient, *ICC*	−1 to 1 (rarely below 0)	Continuous variables	The proportion of the total measurement variability due to variation among individuals Analogous to the kappa statistic, but for continuous variables Higher *ICC* means better agreement
Linear regression, β, c		Continuous variables	Yields a measure of the intercept c and slope β of the regression function
Coefficient of variability, *CV*	0–100%	Continuous variables	The standard deviation expressed as a percentage of the mean value of two sets of paired observations Lower *CV* means better agreement
Average error	0–100%	Continuous variables	The ratio of the mean absolute difference of pairs of measurements to the overall mean value of the measurements Lower average error means better agreement
I_{APSD}	0 to ∞	Continuous variables	The percentage increase in among participant standard deviation due to intra-observer measurement error

identify those individuals who do not have the disease or exposure characteristic of interest. A limitation of the use of sensitivity and specificity is that very few diagnostic tests are inherently dichotomous. Most diagnostic tests are based on the characterization of individuals based on one or more underlying traits, such as blood pressure or serum glucose level. Values for the sensitivity and specificity would vary according to the cut-off level used to separate “diseased” (or exposed) from “undiseased” (or unexposed) individuals. In addition, if measurement error occurs, individuals with true levels of the underlying trait close to the test point are more likely to be misclassified. Since the distribution of underlying traits also determines disease prevalence, sensitivity and specificity can vary from population to population (Brenner and Gefeller 1997).

Example 14.1. Calculation of sensitivity and specificity

		Gold Standard Results		
		Positive	Negative	Total
Study Results	Positive	*a*	*B*	*a* + *b*
	Negative	*c*	*D*	*c* + *d*
		a + *c*	*b* + *d*	

Sensitivity = $a / (a + c)$
Specificity = $d / (b + d)$

Agreement for categorical variables (e.g., X-ray readings by radiologists) is generally reported using variations of the percent agreement and kappa statistics. While overall percent agreement is intuitive and easy to calculate (Example 14.2), it can make agreement look artificially high since there is likely to be considerable agreement between two observers reading negative, or normal, results. An alternative approach is to disregard subjects labeled as negative by both readers to calculate the percent positive agreement (Cicchetti and Feinstein 1990).

Example 14.2. Calculation of percent agreement

		Technician 2	
		Positive	Negative
Technician 1	Positive	*a*	*B*
	Negative	*c*	*D*

Percent agreement = $(a + d)/(a + b + c + d)$ x 100
Percent positive agreement = $a/(a + d + c)$ x 100

Neither overall nor percent positive agreement takes into account the fact that some amount of agreement between two observers will be due to chance alone.

The extent of agreement between two readers beyond that due to chance alone can be estimated by the kappa statistic (Example 14.3) (Agresti 1990; Fleiss 1981; Landis and Koch 1977). In comparisons of more than two categories, a weighted kappa approach allows consideration of the fact that disagreement between some categories may be more serious than disagreement between other categories (Cohen 1968). Like the sensitivity and specificity, variations of the kappa statistic are limited by the fact that most underlying traits are not dichotomous, and different cut-off levels can affect the value of kappa (Maclure and Willett 1987). Interpretation of the kappa statistic should also take into account the fact that kappa can be affected by the prevalence of the condition: for a fixed sensitivity and specificity, kappa tends toward 0 as the prevalence of the condition approaches either zero or one (Thompson and Walter 1988). Additionally, high values of kappa can be obtained if the marginal totals of the contingency table are not balanced (Feinstein and Cicchetti 1990; Maclure and Willett 1987; Thompson and Walter 1988).

Example 14.3. Calculation of kappa for a binary measurement variable

		Technician 2		
		Positive	Negative	**Totals by Technician 1**
Technician 1	Positive	45	5	50 (61.0%)
	Negative	2	30	32 (39.0%)
Totals by Technician 2		47 (57.3%)	35 (42.7%)	82 (100%)

$$\text{Kappa} = \frac{\text{(Proportion Observed Agreement – Proportion Expected Agreement due to Chance)}}{\text{(1.0 – Proportion Expected Agreement due to Chance)}}$$

Proportion Observed Agreement, $P_o = (45+30)/(45+5+2+30) = 0.91$

$$\text{Proportion Expected Agreement due to chance, } P_e = \frac{(50 \times 47)/82 + (32 \times 35)/82}{82} = 0.52$$

$$\text{Kappa} = \frac{P_o - P_e}{(1.0\ - P_e)} = \frac{(0.91 - 0.52)}{(1.0 - 0.52)} = 0.81$$

Common measures of agreement used to assess reliability for continuous measurements (such as blood pressure readings) are the *correlation coefficient*, the *intra-class coefficient,* the *average error*, and the *coefficient of variation* (cf. chapter ▶Analysis of Continuous Covariates and Dose-Effect Analysis of this handbook). *Linear regression* techniques can also be used to check for systematic differences (cf. chapter ▶Regression Methods for Epidemiological Analysis of this handbook).

Although the *Pearson's correlation coefficient*, *r*, is one of the most frequently used measures of agreement in the medical literature, its use is often not appropriate (Altman and Bland 1983; Szklo and Nieto 2000). For one, the correlation coefficient is equally high when both observers read the exact same value and when a systematic difference (bias) exists between observers but the readings vary simultaneously. The value of *r* is also very sensitive to extreme values and the range of values, with a broader distribution of values yielding a higher r. While the Spearman correlation coefficient r_s may be more appropriate to assess the comparability of the rankings of readings and would moreover be less sensitive to outliers, it does not address the main problem of the inability to detect systemic differences between observers.

The *intra-class coefficient* (*ICC*), or the reliability coefficient, estimates the proportion of the total measurement variability due to the variation between individuals (Fleiss 1981). The *ICC* is analogous to the kappa statistic used for categorical variables, and the value can be interpreted in a similar manner. The *ICC* is a true measure of agreement in that it combines information on both the correlation and the systemic differences between readings (Deyo et al. 1991). As with the correlation coefficient, however, the *ICC* is affected by the range of values in the study population.

Other commonly used measures of variability are the *average error* and the *coefficient of variation (CV)*. The average error is the ratio of the mean absolute difference of pairs of measurements to the overall mean value of the measurements. The coefficient of variation is the standard deviation expressed as a percentage of the mean value of sets of replicate observations. In a reliability assessment, the *CV* would be calculated for each pair of observations and then averaged over all pairs of original and replicate measures. A limitation of the *CV* and average error is that both measures may reflect the magnitude of the mean value more than the magnitude of the measurement error (Canner et al. 1983). An alternative measure that has been suggested for assessing variability is the increase in the among-participant standard deviation, the I_{APSD} (Canner et al. 1991). This measure can directly determine the impact of measurement error on the overall among-participant variability for a variable of interest.

Linear regression techniques can estimate systematic differences between readers which are reflected in the slope and intercept of the regression model. One drawback of using regression to assess reliability, however, is that measurement error occurs in both the dependent and independent variables, violating the assumption of an error-free independent variable required for regression (Altman and Bland 1983). However, only under unusual circumstances would measurement error lead to confusing or uninformative results.

14.3.7 Planning Data Management

The management of data in a large epidemiological study can be a formidable task. The sheer volume of data for a sizeable study with extended follow-up can become quite overwhelming, as illustrated by the following example. If 100 data elements are to be collected for each participant in a cohort study with 100,000 participants, the data collected at the end of each data collection cycle are comprised of 10^6 distinct data elements. Let us say that in order to update exposure and outcome information, data are to be collected yearly for each participant for 10 years. This increases the amount of data being collected by an order of magnitude, to 10^7distinct pieces of data. Superimposing on this volume of data the errors that can occur during data recording, transcription, and transfer of data to an electronic medium, it is easy to see how data quality can be compromised without careful planning of how data are to be managed. The potential magnitude of the task of data correction is also clear.

The first step in planning a data management system is to define what data will be collected and how often it will be collected, keeping in mind that as the volume of data increases, ensuring data accuracy becomes more difficult. In order to further minimize the amount of unnecessary data collected, the chosen data variables should be prioritized, and a "tolerance" of error established for each data field. For example, it might be decided that all values of crucial data variables (e.g., disease outcome) should be checked against written questionnaires, but auditing a random sample of questionnaires is sufficient for other fields.

The next key step in data management planning is to define essential identifying information for the study data. *Identifiers*, generally known as *key*, *header*, or *ID* fields, are fields that allow each form to be uniquely identified and correctly related to other forms (Hosking and Rochon 1982; Hosking et al. 1995). Study identifiers are usually located in a standard header section of the form. Entry of an incorrect number into one of these fields can cause the entire data record to be processed incorrectly.

Most studies require at least four types of identifiers: study identifiers, participant identifiers, form-type identifiers, and timepoint identifiers. Depending on the study, other identifiers (e.g., family identifiers) might also be necessary. *Study identifiers* designate the sponsor, study, protocol, or substudy. *Participant identifiers* uniquely identify the study participant. In general, a study-created participant identifier is preferable to a natural identifier such as participant name or social security number, especially in a climate increasingly concerned with participant confidentiality. It is useful to encode information about participant characteristics (such as a field site code) into the participant identifier since this allows later classification of participants by their identifiers alone. *Form identifiers* identify a particular questionnaire and often take the format of a two- or three-character abbreviation. Form identifiers in longitudinal studies should be planned so that multiple versions of each form can be accommodated. Adding a −1 at the end of the form identifier, for example, allows for future versions to end with the suffix −2 or higher. *Item identifiers* form the last group of identifiers and are assigned to each question on a form. While item identifiers bearing a one-to-one relationship to database fields might be useful for data analysis, this can become confusing if the study forms or database are revised. Data management systems that track the relationship between each database field and a corresponding item number in each form version provide a useful alternative.

Once data identifiers have been selected, general data management considerations that need to be addressed including identifying how data will be entered (electronically vs. manually); who will do the data entry; what software will be used; what types of edits will occur during data entry; how queries will be generated, communicated, and resolved; how suspicious values will be treated; and how corrections will be implemented and documented. The remainder of this section will be devoted to these considerations.

Table 14.3 Overview of the traditional data management process (Adapted from DuChene et al. (1986))

Steps in data processing	
Data collection and mailing	Complete forms at clinic/in field Visually review form while participant is in clinic (visual editing) Mail original copy to coordinating center, keep copy at clinic Create standard packing list for mailing
Receipt and conversion to electronic format	Receive forms at coordinating center Acknowledge receipt of forms from clinic using postcard or electronic mail Log receipt of forms into computer (including form number, ID code, date completed, date received, and unique log number) Key the form. Verify keying
Forms processing, posting, and backup	Process form through an edit program that checks type and range of each field, as well as internal consistency of form Generate computer edit report Check edit report and initiate appropriate error correction procedures Back up edited forms Post forms to master file. Back up master file File form
Clearance and archiving	Run further checks on data to ensure that posted data are consistent with other data on file Review edit reports that result from checks and initiate appropriate error correction procedures Document master file contents and prepare file for archiving

14.3.7.1 Design of a Data Management System

In a multicenter study, data are typically collected at various field centers and then sent to a coordinating center for processing, storage, and analysis. Table 14.3 provides an overview of the steps involved in data management. Newer approaches to data management (e.g., Web-based systems) are increasingly being used (Aanensen et al. 2009; Kwok 2009), but these approaches rely heavily on specialized automated systems for data collection, entry, and auditing. Since logistical barriers of cost, lack of expertise, and low computer literacy may render these systems impractical for some investigators, this chapter focuses on a more traditional approach to data management. Many of the underlying principles remain relevant to the newer data management approaches, which are addressed at the end of the section.

14.3.7.2 Data Recording and Visual Editing

The measurement and recording of data from study participants usually occurs at the field center, and initial data checks are generally conducted by field staff personnel. Field center interviewers or technicians should check data for consistency as it is being collected, while the study participant is available to clarify any immediate discrepancies, errors, or out-of-range characteristics.

For technical measurements, an independent review of samples by two or more readers should be performed on all or a subset of samples. This allows later assessment of validity and enables investigators to track down sources of error. On completion of the data form, field center staff should perform a routine review of forms to establish that the questionnaire is complete, that skip patterns have been followed, and that the data values appear reasonable. If routine review of the form does not identify any unusual data, the form can be processed further. Including an indication of who reviewed the form will facilitate later examination of the editing process.

14.3.7.3 Data Entry

Almost universally, epidemiological data are entered into electronic databases for storage and analysis. The processing, storage, and analysis of study data usually occurs at the data coordinating center. Errors that can occur during the processing and storage of data include keying errors, inaccurate data transcription, and programming errors (Arts et al. 2002). In the Hypertension Prevention Trial, key error was found to be the major source of data entry error, with 5.2/1,000 errors out of an overall error rate of 6.9 errors per 1,000 data items being key errors (Prud'homme et al. 1989).

Most automated data entry systems allow a variety of mechanisms for checking data. As data entry is initiated, form identifiers are checked for validity and consistency. *Range checks* during data entry can be used to electronically limit the data type or the range of possible values at entry. For example, date fields can be programmed to accept only valid dates, or table look-up systems can restrict the values of categorical data with a limited number of possible values. For continuous data, many studies use normal population ranges of a variable to flag outliers. While programmed range checks are a useful tool, retaining some flexibility to correct errors at the time of data entry is important since too many restrictions on modifying data at entry can lead to a higher error rate (Crombie and Irving 1986).

Data accuracy can also be improved by the use of *double data entry*. The independent keying of data twice, however, does not prevent all types of error. Examples of errors that would not be reduced by the double entry of data include errors in transcription or misinterpretation of data in the same way by two data entry operators.

If the data are to be manually entered, personnel should be masked regarding exposure or outcome status (depending on the study design) to prevent the possibility of observer bias. Additionally, the electronic database should have a provision to indicate who entered the data to allow for later review of data entry performance.

The use of electronic technology for data entry as an alternative to manual data entry is gaining in popularity. Software is available for scanning forms directly into an electronic database using optical character recognition (OCR). The accuracy of scanning software is quite variable, however, and a process to check scanned data should be in place. In general, OCR is better suited to numeric or check-box responses than to hand-printed characters. Another method of electronic data entry is computer-assisted data collection (CADC), whereby interviewers directly enter participant responses into a computer file. This technology completely circumvents the need for transferring data from paper to an electronic medium, thus eliminating errors associated with this process. A CADC system can automatically enforce skip rules, require completion of required fields, and flag suspicious values for correction while the study subject is still present. Since errors in CADC data cannot be compared later with a paper form, however, these systems need to include as many ways of checking data accuracy at entry as possible. One way to allow for examination of inconsistent values is to tape record interviews while data collection is occurring.

In a pilot study of CADC, five study staff members with no prior experience using a CADC system were trained and asked to administer both CADC and paper-based interviewers to 16 study participants. All five staff members preferred the CADC system, indicating faster and more accurate data entry and less likelihood of erroneously skipping an item. Ten of the 16 pilot study participants had no preference between paper and CADC and six preferred CADC. Although the median time for data collection at the reception, examination and interview stations was slightly longer for CADC than for paper interviews, the CADC data are already partially edited and in machine-readable format, whereas data from the paper forms still had to be edited and keyed. The percentage of suspicious data values was similar for each method, but 21 of the 25 suspicious data values were identified and corrected at the time of collection using CADC, compared to 1 out of 23 suspicious values corrected with the paper system (Christiansen et al. 1990).

Other recent methods of data collection for epidemiological studies include the use of electronic mail ("e-mail") (Kiesler and Sproull 1986; McMahon et al. 2003; Paolo et al. 2000), personal digital assistants (PDAs) or mobile phones (Aanensen et al. 2009; Kwok 2009), or other Internet-based surveys (Baer et al. 2002; Blackmore et al. 2003; Rhodes et al. 2003; Silver et al. 2002; Turpin et al. 2003). E-mail questionnaires have been reported to have a faster rate of return and more thorough completion of returned questionnaires, but response rates have generally been lower than for mail questionnaires.

The basic process for Internet-based, or Web-based, data collection is the translation of the study questionnaire into an Internet language (HTML or hypertext markup language) and posting of the questionnaire onto the World Wide Web. Respondents then complete the survey using a point and click interface. The survey is generally visually and functionally similar to traditional surveys.

Web-based data collection provides several advantages over paper form data collection. For one, researchers can access populations that previously might have been unreachable due to geographical or cultural boundaries. Use of the web may also speed up the time of data collection since no testing site or appointment scheduling is necessary, and the need for data entry by study personnel is eliminated. Web-based systems can also minimize the variation due to differences in survey administration, interviewer interpretation, and entry of data. Since complicated branch and skip patterns can be programmed into the survey, the amount of interviewer or respondent attention necessary is reduced. Costs can drop dramatically with the use of web-based data collection, as there is no need for printing, mailing, and data collection personnel. Web-based surveys also provide a greater degree of anonymity for the collection of sensitive personal information (Baer et al. 2002).

It is important to realize, however, that depending on the situation, some advantages of web-based systems can become disadvantages. In some populations or countries, for example, the cost of printing, mailing, and administering a paper questionnaire might be considerably less than the cost of setting up a web-based system and providing training and access to study participants. Other disadvantages of web-based data collection include the possibility of selection bias when choosing a study population and security problems during data transmission. The issue of computer users being unrepresentative of the general population can be overcome to some extent by providing Internet access to a randomly sampled study population (Silver et al. 2002). Literacy or language barriers, however, may still prove to be an issue.

Incorporating strict security measures in an electronic data entry system is crucial to maintaining data confidentiality and can require considerable time and monetary resources. In some instances, it might be possible to provide a quick solution to this problem by linking the survey security to an existing high-security system, such as a university network.

While the use of web-based systems is a promising avenue of data collection for studies, such systems require considerable expertise for adequate setup. Often, initial versions of web-based questionnaires present frustrating technical problems to users and may require several iterations before a working system is in place. Web-based systems may be inappropriate for populations that are not computer literate. Additionally, data entry errors by users can still occur. Ethical concerns which arise in the study still need to be addressed and may be more difficult in the context of web-based data collection. For example, the investigators still bear responsibility for verifying informed consent or for providing local targeted support in case the respondent needs a referral as a result of the research. For these reasons, the pretesting of data instruments, postentry error-checking, and other forms of data quality control described in this chapter are as crucial for more technologically advanced data collection as they are for more traditional forms of data collection.

14.3.7.4 Data Audits

Once data have been entered, they must be submitted to further accuracy checks. One method of assessing data accuracy is to perform a series of consistency checks, such as ensuring that the date of birth and age of a participant are in agreement. Reviewing samples that fall outside some number of standard deviations of the mean is a sensible alternative way to check data. More formal statistical methods for detecting outliers can also be used (Barnett and Lewis 1994; Vardeman and Jobe 1999). The importance of using range checks is illustrated by a simulation in which different rates of entry error were introduced into a constructed dataset and simple range checks were used to identify and correct outliers. Even with a random entry error rate as high as 20%, population means remained very similar after the correction of unusual values, regardless of study sample size (Day et al. 1998). Error rates similar to those achieved with double data entry were achievable when extensive logic checking of fields was incorporated (Mullooly 1990; Neaton et al. 1990).

In instances where an unusual value is detected, a data quality query should be generated either manually or automatically. During study planning, a system for reporting and responding to such queries needs be conceptualized, along with the designation of individuals responsible for checking and responding to questions. The automatic generation of regular quality control reports including summary statistics such as the number of queries by form and data field, or the percentage of error-free forms, can aid the systematic processing of data. Section 14.4 of this chapter addresses the processing and resolution of error queries in more detail.

Comparing the number of forms that are edited using automated checks at the data-processing center to the number of forms recorded in the batch sent by the field center allows the identification of forms that are lost during keying. Additionally, a random sample of data forms should be compared to the electronic data submitted to check accuracy of data entry.

Once routine edits have been completed, the data form can be posted directly to a *master file* for smaller studies, or to a *distributor file*, for larger studies. In the Multiple Risk Factor Intervention Trial, the edited form was transferred to a distributor file, which held all the forms that were edited in a day. At the end of the day, forms held in the distributor file were transferred to one or more *transaction files*, which served as temporary storage until the next scheduled update of the master file. The use of transaction files allowed investigators the flexibility to resolve discrepancies before the data were added to the master file. Transaction files were generally copied to daily backup tapes so that data could be retrieved to the time of the last backup in case of processing errors, machine failure, or other accidents (DuChene et al. 1986).

14.3.7.5 Forms Posting

In general, it is best to keep the interval between data collection and entry as short as possible. If it is possible to process forms as they are generated, this is preferable (Meinert and Tonascia 1986). If batch processing is found to be more convenient, the scheduled time between subsequent postings of information from the study transaction files (raw data) to the master file should not be longer than 2 weeks. During forms posting, data fields from the transaction files are copied to the location in the master file(s) specified in the *data dictionary* (a database of information used to edit, document, and control the processing of forms through the computer system). The data management system should be programmed to reject the form if errors are detected in the data identifiers, or if data are found to already exist in the master file (unless the form to be entered is a correction form). Personnel at the data coordinating center can then review and resolve discrepancies in rejected forms. If fields need to be modified in the master file, changes should be explained and documented in the electronic file as well as on paper.

14.3.7.6 Backup of Raw Data

Once the forms have been posted to the master file, all transaction files containing the posted forms should be copied on to a tape or other electronic medium such as compact disc (CD), digital video display (DVD), or external hard drive, and stored offsite in a secure facility. In the event of a major system failure or destruction of the master file (in a building fire, for instance), the offsite copy will allow recreation of the master file.

14.3.7.7 Clearance

After the data are posted to a master file, computer edits of the master file allow consistency checks between fields on different forms. For example, an individual's height should remain constant over forms. It is informative to flag inadmissible values, as well as unlikely values. Additional within form checks can also be performed at this time.

14.3.7.8 Archiving

When within-form verification and across-form clearance are complete, and data on the master file are finalized, the master file should be copied on to at least two tapes and stored offsite. These tapes should be read regularly to check for deterioration. If a backup tape cannot be read, a new copy should be made.

14.3.7.9 Networks

With the emergence of networks or cohort consortia in large-scale epidemiological studies, data management becomes an increasingly complex issue. The need for standardization among study cohorts within a network requires an additional level of organizational coordination of data management and quality assurance practices. Some consortia establish steering committees to oversee quality assurance issues. Steering committees can serve to guarantee efficient and accurate data management while minimizing and streamlining administration. These activities are likely to involve harmonization of data that have been collected and/or coded in different ways. Reconciliation of disparate datasets may prove quite challenging. For more information on quality assurance in network settings, see Khoury et al. (2010).

14.3.8 Quality Assurance Committee

The most carefully designed quality assurance program cannot function efficiently without the assignation of responsibilities for various quality monitoring tasks to specific individuals and the existence of effective communication channels between study personnel. In many large studies, a quality assurance committee is formed to oversee the quality of data collection (Knatterud et al. 1998; The Chemical Manufacturers Association's Epidemiology Task Force 1991; US Environmental Protection Agency (EPA) 1989). The quality assurance committee addresses quality issues throughout the life of the study, from protocol development to the responsible archiving of data. The quality assurance committee is also responsible for reviewing study compliance with written quality assurance/control procedures and for evaluating interim analyses. For large studies, a data monitoring committee made up of external quality assurance auditors supportive of the protocol objectives and study design might be warranted (Fleming 1993; Khoury et al. 2010).

14.3.9 Communications

The effective resolution of study quality issues is highly dependent on the quality of communications between study personnel. Many of the quality assurance mechanisms already described in the chapter *contribute* directly to improved communication. Examples include the training of personnel and the definition of standard operating procedures. Other quality assurance mechanisms *depend* critically on communication for their implementation. In order for queries to be resolved effectively, study personnel need to know who to submit queries to and how these queries should be submitted. Structures for transmitting resolved data queries back to data entry personnel are also needed. The scheduling of regular meetings between study personnel is crucial for maintaining study communications. Emphasizing the rationale for quality control and the need for wholehearted support for quality control measures is important since quality control measures will fail if they are perceived as nit-picky and burdensome (Cooper 1986). One or more individuals should be designated responsible for preparing and disseminating the minutes of study meetings. More generally, communication structures should be in place to communicate the intent, conduct, results, and interpretations of the study to study personnel, study participants, and the scientific community. In certain situations, other parties that might need to be informed of study results include health care providers, policy makers, or the media.

In network or cohort consortia settings, communication among study groups is particularly important with regard to authorship and publications. Issues of authorship and publication policies should be established early to avoid confusion and dissent that may arise later.

14.3.10 Cost of Quality Assurance

Clearly, the implementation of quality assurance and quality control measures add to the cost of a study. While some expenses, such as the cost of routine data editing or the rechecking of statistical analyses may be impossible to estimate, cost information can be projected for other aspects of quality assurance, such as training, site visits, and external quality control programs (Knatterud et al. 1998). Considering the cost of various quality control measures early in the planning process allows for development of a realistic and feasible program that is more likely to be executed. Priorities for data quality should be set at this time. While certain aspects of data quality should not be sacrificed regardless of the expense, a compromise might be possible in other instances. For example, a costly, time-consuming measure of exposure might be collected for a subsample of study participants and this information can be used to validate a cheaper exposure measurement used for all study participants.

14.3.11 Ethical Considerations

Ethical considerations are perhaps the most important set of considerations in a study (for a general discussion, see chapter ▶Ethical Aspects of Epidemiological Research of this handbook). Epidemiological research should never lose sight of the fact that data are derived from human beings. Studies such as the Tuskegee Syphilis Trial (US Department of Health, Education, and Welfare (DHEW) 1973) which followed the progress of untreated syphilis in black men even after effective treatment was available may now seem shocking, but it is well to keep in mind that throughout most of the trial, the investigators did not find their research particularly objectionable. The thorough consideration of ethical issues raised by a study (mandated by law in most countries) will hopefully prevent a future generation of scientists from looking back at present-day trials with regret.

The human subjects section of the protocol must describe whether the study protocol imposes any physical or psychological risk to the participants. Potential benefits of the study should also be noted, with an explanation of whether benefits will be accrued by study participants themselves or whether the study is expected to benefit others in the future. The cost-to-benefit ratio should be weighed and discussed. Studies that involve primary data collection generally need to obtain informed consent from study participants. Consent forms should include, at a minimum, contact information for personnel available to answer questions about the research, the purpose of the study, eligibility requirements, the expected duration of participation, possible harm that the subjects could incur, expected benefits to subjects or to others, information on the voluntary nature of participation, and a statement indicating the right to withdraw from the study at any time (The Chemical Manufacturers Association's Epidemiology Task Force 1991). The study eligibility criteria are also subject to ethical considerations, both in terms of inclusions (different racial/ethnic groups and both sexes should be adequately represented) and exclusions (special justification is needed for study of vulnerable groups, such as pregnant women, children, or incarcerated individuals). Adequate provisions for maintaining data confidentiality and the privacy of individuals should be described. For example, investigators might plan to store hard copies of sensitive data in locked cabinets with limited access and remove personal identifiers from datasets used for analysis. Automated data management systems should have password control, users should be logged out after a period of inactivity, and the copying of data should be discouraged (Wyatt 1995).

Some of the most pressing ethical challenges to date arise from large-scale studies which rely on aggregating databases of biological specimens or "biobanks." Specimens included in biobanks may have been collected in the past with participant consent that does not meet today's standards of informed consent. In fact, the specimens may be used for assays and analyses that were not anticipated at the time of study initiation. Furthermore, participant consent may not apply to secondary uses of their data for future studies as yet unanticipated or for studies conducted by other investigators. Several recommendations have been made to address the ethical challenges of data sharing and aggregated databases, including ensuring the existence of data security mechanisms to protect participant privacy, accounting for the potential incorporation of data into aggregated databases when obtaining informed consent, and establishing data sharing rules (Karp et al. 2008).

14.4 Quality Considerations During Study Conduct

Before data collection is initiated, all data collection procedures should be reviewed and approved by the lead investigators. Data forms and equipment should have been tested and certified ready for use.

If rigorous quality assurance procedures have been planned prior to study initiation, quality control activities during study conduct mainly consist of the implementation of these procedures. The study protocol should be followed, personnel should be trained according to established standard procedures, and data collection should proceed with all quality assurances in place. Any deviation from a given standard operating procedure should be authorized by the steering committee.

The importance of periodic examination of data by study investigators, data coordinators, and data entry personnel while the data are being collected cannot be overstated. Examination of data trends by center, over time, or by technician (for example), can identify flaws in data collection early on. Even simple plots and graphs of data can identify sources of error. When data errors are identified, steps should be taken to correct the data in a timely manner. In some cases, statistical adjustment can be used to correct data drift. When this is not possible, data might have to be thrown out or completely reprocessed. In order to generate a written audit trail of data, any changes made in the data should be documented.

14.4.1 Training and Certification

The importance of training and certifying all study personnel has already been underlined in Sect. 14.3.3 “Personnel: Training and Certification.” While many study investigators are aware of the need for standardized operating procedures, information regarding these procedures is often lacking in study descriptions. While 244 original research articles in three emergency medical journals (1989–1993) described data collection by means of chart review, only 18% mentioned training of abstractors and periodic abstracter monitoring was reported in a mere 3% of these articles (Gilbert et al. 1996).

Detailed practical guidelines for training and quality control management for study interviewers, data abstractors, and biomedical technicians are available in the literature (Edwards et al. 1994; Fowler and Mangione 1986, 1990; Reisch et al. 2003). This section summarizes some of the main considerations.

14.4.1.1 Training

Training procedures should ideally involve all staff and procedures. While centralized training of all study personnel might be desirable in terms of increasing the comparability of data collection between sites and allowing study personnel from different sites to interact with each other, the expense of bringing personnel to a central training site for all their training can be considerable. Additionally, site-specific questions might arise that cannot be adequately addressed during centralized training. An optimum strategy might be to use both types of training. Table 14.4 provides an overview of the training process.

Table 14.4 Overview of training (Adapted from Reisch et al. (2003))

Steps in data processing	
Training manual	Educational training manual is sent to all sites for review The training manual consists of some or all of the following: a study overview, information on the relevant procedure, quality assessment procedures, data forms with instructions (e.g., for abstraction or interview), quick reference sheet for all variables, glossary of terms, standardized training examples, and relevant articles from the literature
Standardized training examples	Training examples should be prepared for key study variables. Study personnel might be asked to note blood pressure measurements from a training tape, for instance
Individual orientation	Two or more individual orientation sessions should be arranged with the onsite data collection team and with the lead study coordinator and/or study investigator. Additional sessions can be scheduled at the discretion of the site coordinators
Double-review of initial data	The first few examples of data collected (by chart abstraction, interview, or a biomedical procedure) should be repeated by a more experienced member of the data collection team. Discrepancies can then be reviewed Queries should be entered into an audit form and sent to the lead study coordinator to assist with later tracking of problematic data
Regular double-review	Performing regular double review for a small sample of data (e.g., once a month) can prevent data drift over time. Review of data at a later time is facilitated by audio or video taping of interviews or biomedical procedures
Regular conference calls/ meetings of field staff	Regular study conference calls can include a training component if examples of data collection problems are brought up for discussion during each call. An updated decision log containing a summary of discussions held and decisions made during these conference calls can be distributed among study personnel
Regular site visits	Review of data collection procedures during site visits by the lead study coordinator and/or lead investigator
Retraining	Retraining study personnel might be necessary if substantial time has passed since initial training, a systematic bias in data is detected, or the study protocol changes

14.4.1.2 Certification

Following initial training, study personnel should be certified to perform specific procedures. Regular retraining is desirable to prevent data drift. Retraining might also be necessary if a specific study technician is found to be introducing a systematic error into the data, or if the study protocol changes. Any retraining should be accompanied by recertification.

While the interval between retraining and certification varies from study to study, the Atherosclerosis Risk in Communities study (ARIC) used a 90-day interval since a 6-month interval was found to allow too much drift to recognize and correct digit preference. More timely feedback was also needed in the Cardiovascular Health Study (CHS) (Joel Hill, personal communication 2003).

14.4.2 Maintenance and Calibration of Equipment

Study equipment should be inspected and calibrated at regular intervals in accordance with the study protocol. In the event of equipment breakdown, equipment may need to be replaced. If the new equipment is similar to the equipment already being used, then calibration before use is sufficient. When replacement of existing equipment is desirable because a new model or instrument is more accurate or efficient than the existing equipment, data should be collected using both the old and the new instrument for a defined period of time, so that comparability of measurements can be established.

14.4.3 Implementing Data Management

The data management process has already been described in detail in Sect. 14.3.7. During study conduct, the planned data management system is implemented and refined as necessary.

14.4.3.1 Tracking and Monitoring of Data

The effective tracking and monitoring of data as data collection is in progress is essential to the timely detection and correction of errors. Monitoring should occur for subject accrual, data acquisition, and data quality. Automated tracking systems can greatly assist this process and have been used successfully in epidemiological studies as early as 1981 (McQuade et al. 1983). Data that are collected by hand should be recorded directly, promptly, and legibly in ink. Four different types of monitoring are recommended: *proactive efforts* to improve data, *observation of data collection*, *review of computer-generated checks and summary reports*, and *examination of data*.

When possible, data quality should be improved by *proactive efforts*. Automated reminders of when patients are due for study visits for time-dependent variables (e.g., levels of an exposure biomarker) can prevent the collection of data that is later deemed of poor quality or unusable. Target dates for follow-up visits can be defined by the participant's entry date rather than the date of the last visit, in order to prevent scheduling deviations from carrying over to future visits.

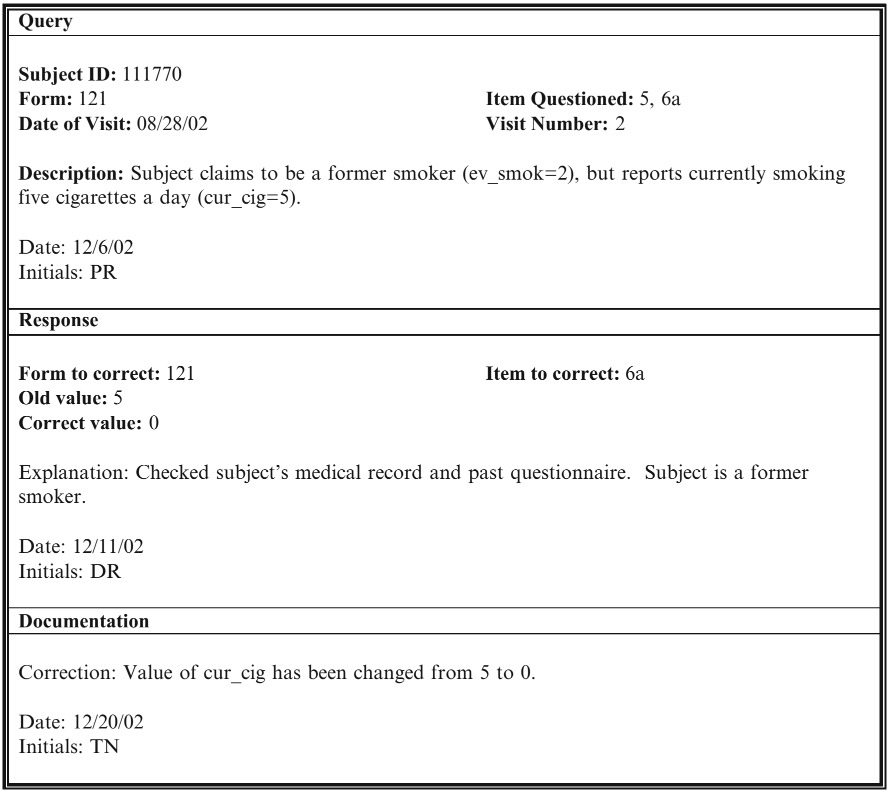
Direct or indirect *observation of data collection* can also identify errors in a timely manner. An unobtrusive way to monitor interviewers for delivery and adherence to protocol is to audio-record interviews. Measurement techniques for biomedical or laboratory technicians can either be video recorded or directly observed by senior technicians or other qualified study personnel.

Regular *review of computer-generated queries and summary reports* of data quality can alert the investigators to a variety of data errors, including participant ineligibility, data outside the expected range, and variation in data quality by data field, site, or technician. Active examination of data during collection is crucial. Summary statistics and plots of data by technician, site, or time can identify unusual trends. For example, an examination of data from the Hypertension Prevention Trial revealed that nearly 29% of the baseline systolic blood pressure readings from one clinic ended in the digit 2. This could be traced to measurements made by one technician, who recorded a number ending in the digit 2 for over 60% measurements (Canner et al. 1991). When a data collection flaw is identified, further error should be prevented by tracking down the source of the problem and taking corrective action.

Keying errors may be identified by periodic audits of the database against source documents. Rather than check all the data, a random sample of data fields can be selected to check for keying errors. When creating the test sample, it is important to ensure that a broad cross-section of data is included (e.g., both numerical and character fields should be checked). One method for sampling a variety of fields is to choose a random subsample of forms and look at all fields within those forms.

14.4.3.2 Corrective Actions

Moving back to the datascope for a moment, we recall that the identification of data errors is only the first step in data quality management. In order to reach the ultimate goal of valid data, these errors need to be corrected. The process for revising data should be as systemized and well documented as the process for locating errors. While the routine correction of careless mistakes while data entry is in progress need not be reported, data errors that are identified after initial data entry should not be changed by data entry staff until the query has been checked. A paper trail should be initiated for each problem, with the initial query describing the problem, and the date it was detected (Fig. 14.2). The individual(s) responsible for query resolution should then investigate the query and provide a response explaining why the problem occurred. Finally, the query documentation should indicate how and when the problem was resolved. If data from a form are found to be incorrect, they should be identified as incorrect rather than erased, and the correct values



Query

Subject ID: 111770
Form: 121 **Item Questioned:** 5, 6a
Date of Visit: 08/28/02 **Visit Number:** 2

Description: Subject claims to be a former smoker (ev_smok=2), but reports currently smoking five cigarettes a day (cur_cig=5).

Date: 12/6/02
Initials: PR

Response

Form to correct: 121 **Item to correct:** 6a
Old value: 5
Correct value: 0

Explanation: Checked subject's medical record and past questionnaire. Subject is a former smoker.

Date: 12/11/02
Initials: DR

Documentation

Correction: Value of cur_cig has been changed from 5 to 0.

Date: 12/20/02
Initials: TN

Fig. 14.2 Example of a data query form

should be inserted (Knatterud et al. 1998). In some cases, unusual values will be confirmed to be correct in which case they should be retained in the database with documentation.

Occasionally, errors identified during study conduct may lead to changes in the survey instrument or other study equipment. In such cases, it is crucial that the version of the form or equipment used to collect data is recorded in the database. If a new data check is added, either as a result of a query or as an additional precaution, old values in the database should be edited using the new rules in order to keep data consistent.

Tracking the time taken for corrective actions allows areas of delay to be identified and resolved for future queries. In most longitudinal studies, data are analyzed while data collection is still in progress. In such instances, one might want to exclude data that are under query from the master database until the problem is resolved. The inclusion of a "status" field for data would allow investigators to check whether values were acceptable or unacceptable (Gassman et al. 1995).

14.4.4 Site Visits

For multicenter studies, site visits to observe operations allow greater understanding of site-specific data collection issues and provide an opportunity to recognize and correct faulty systems (Gassman et al. 1995; Knatterud et al. 1998; Prud'homme et al. 1989). Scheduling site visits is recommended shortly after initiation of patient recruitment and when the data collection at the site is drawing to a close. Additional site visits should be scheduled for long-term studies.

The size of the site visit team can vary and is dictated by the nature and purpose of the visit. A typical site visit team might include the study principal investigator (or representative), the director of another field site, the data coordinating center director, the study project officer, and selected resource personnel. During the site visit, the site visit team would meet with the director and staff of the unit and hold private conversations with key support personnel. The site visit should include a thorough review of staffing requirements, recruiting, training and certification, and communication structures. Site visitors also have a chance to observe data collection, check data management, and review data quality monitoring. Specific activities might include observation of whether field technicians follow the study protocol, inspection of study records and documents storage, and review of the operation and maintenance of local data systems. Following the site visit, the leader of the site visit team should prepare a written report of the visit based on input from the entire team. The site visit report should describe any systematic errors that were identified in data collection and provide recommendations on how to rectify the situation. A formal response to the report should be prepared by the staff at the study site.

14.5 Quality Considerations After Data Collection

Once data collection for the study is complete, the task of analyzing and interpreting the data begins. The study investigator should yet again consult the datascope to check for possible biases and errors that need to be resolved in order to form a clear picture of the relationship under study.

14.5.1 Reporting Response Rate

If individuals who agreed to participate in the study were different in some important way from non-respondents, the study results could well be biased. For example, non-respondents to questionnaires might be of poorer health or more likely to be smokers than respondents (Shahar et al. 1996). Studies that have followed respondents and non-respondents to questionnaires have reported that non-respondents have a significantly higher risk of myocardial infarction, cancer mortality, and all-cause mortality (Bisgard et al. 1994; Heilbrun et al. 1991).

Calculating the study *response rate* gives a first indication of whether the investigator should be concerned about possible bias in the results. Generally, the higher the study response rate, the less need to worry about selection bias affecting the results. The simplest approach to response rate calculation is to divide the number of surveys received by the number of surveys sent. However, this does not account for factors that can affect the response rate such as undelivered questionnaires, ineligibility of subjects who completed questionnaires, or substitution of the intended recipient with another subject. Typically, the numerator and denominator of the response rate are adjusted to reflect such factors. Standard definitions and methods to calculate survey response rates are provided by the American Association for Public Opinion Research (2000) or the Council of American Survey Research Organizations (CASRO).

For cohort studies, the simplest way to estimate the *follow-up rate* is to divide the number of participants seen at the last visit by the number of participants initially enrolled. Again, different assumptions about individuals lost to follow-up yield different numbers for the follow-up rate.

Since different methods of calculating the response rate might be appropriate for different studies, the choice of the response rate formula is less critical than the identification and reporting of all the elements that enter the calculation (Table 14.5).

In general, response rates to questionnaires have been decreasing in the United States, and perhaps elsewhere (Kessler et al. 1995; Steeh 1981). Data from a nationwide survey in the United States (the Behavioral Risk Factor Surveillance System, BRFSS) indicate that response rates from random digit dialing have declined from a median of 68.4% in 1995 to a median of 55.2% in 1999 (Centers for Disease Control and Prevention (CDC) 1999). A review of 82 case-control studies published in the *American Journal of Epidemiology* (1988–1990), *Epidemiology* (1997–1999), and *Cancer Epidemiology, Biomarkers and Prevention* (1997–1999) reported a 0.2% and 0.44% decrease in reported response per year for cases and controls, respectively (Olson 2001). The same article reported an average response rate of 76.1% for cases and 71.5% for controls. A review of 321 distinct mail surveys published in a broader spectrum of United States journals in 1991 reported an average survey response rate of 62% (Asch et al. 1997).

Regardless of the exact value of the response rate, the characterization of non-respondents is crucial in order to assess whether a bias is present and, if it is, how the results of the analysis might be affected. Clearly, describing the non-respondents becomes more important when a study has a low response rate. Whenever possible, a brief survey should be administered to non-respondents to collect limited data for comparison with respondents. Otherwise, assessing available data on demographics, exposure, or outcome will allow some assessment of possible bias.

14.5.2 Analysis

Before proceeding to analysis, the study data should be tested rigorously to check for *residual errors* that remain after all data processing and routine quality assurance

Table 14.5 Reporting outcomes of recruiting respondents in case-control studies in a study of thyroid cancer in Western Washington (Adapted from Olson et al. (2002))

Units selected from sampling frame	Number
Random digit dialing screening phase	
Total	6, 741
Ineligible sampling unit	
Total	3, 589
Business, fax, government	1, 937
Non-working numbers	1, 436
Institution, group quarters, dataline	216
Unable to determine eligibility	
Total	431
Unknown if residential	274
Residential, unknown if individual eligible	157
Answering machine on all attempts	56
Refusal to answer questions on eligibility	76
Other (language barrier)	25
Respondent not eligible	
Total	1, 983
Age	1, 749
County	216
Language	18
Respondent screened and eligible, total	738
In-person interviews of eligible women	
Total	738
Unable to determine eligibility	0
Respondent not eligible	
Total	1
Prior thyroid cancer	1
Respondent screened and eligible	
Total	737
Not interviewed (refused)	163
Interviewed	574

activities are complete. Range checks provide one way to examine whether the data seem reasonable. Simple queries, for example, checking that the recorded age in years is consistent with the date of interview minus the recorded date of birth, can also help to detect errors.

Once the investigator feels confident that there are no obvious flaws in the data, the next step is to understand the data by conducting exploratory data analysis using univariate and bivariate summaries, as well as plots and graphs of the data. More complex exploratory analysis of the data should be guided by the data. If assumptions implicit in the planned analysis methods are violated, alternative statistical methods must be considered. Appropriate and careful statistical analysis is integral to good epidemiological practice. A description of basic methods of analysis for epidemiological study designs can be found in Part III (Statistical

Methods in Epidemiology) of this handbook and in most intermediate textbooks of epidemiology (Rothman and Greenland 1998; Szklo and Nieto 2000). Some of the key issues underlying the analysis of case-control and cohort studies are summarized in chapters ▸Cohort Studies and ▸Case-Control Studies of this handbook and in a two-volume series published by the International Agency for Cancer Research (Breslow and Day 1980, 1987). The finer points of analysis, however, are study specific. For this reason, it is crucial that data analysis be conducted by personnel with the necessary training and experience in statistical methods.

Once data analysis is complete, ways to check the analysis include independently reproducing the tabulations and statistical calculations from the original data and checking different tables for consistency of the denominators. All data reduction and statistical procedures should be documented to facilitate review at a later date.

The results of any study are associated with some degree of uncertainty. To the extent possible, these uncertainties should be quantified and accounted for or, at the minimum, characterized quantitatively. In an analysis of risk factors for coronary disease in the Framingham Heart Study, estimates of risk increased for factors measured with substantial error after correction for uncertainty (e.g., serum cholesterol), whereas risk estimates tended to remain unchanged for risk factors with little or no error, such as body mass index (Rosner et al. 1992).

The analysis of study data is followed by the task of interpreting the study results. An observed association might be due to statistical artifact, due to bias or confounding, or be truly causal. The use of statistical significance alone to guide inference is not recommended (Goodman 1999a, b). In a study exploring 100 associations, five would be statistically significant at the $\alpha = 0.05$ level by chance alone. Moreover, an association might be confounded by one or more variables or could be biased due to systematic flaws in the design or conduct of the study.

The capability of generating very large volumes of data from microarrays and other new technologies has led to new analytical challenges and the emergence of the field of bioinformatics. For example, genome-wide association studies (GWAS) currently explore associations of hundreds of thousands of single-nucleotide polymorphisms (SNPs) with disease status. Soon, researchers will be exploring the actual genetic sequences in people with and without disease, leading to even more massive datasets and analytical challenges. Not only is there a need for multiple comparisons adjustment but also for planning studies with sufficient power to detect small effects (Khoury et al. 2010). The planning of such studies necessarily involves a team that includes genetic epidemiologists, bioinformaticians, biostatisticians, and geneticists and clinicians.

Following adequate consideration of chance, confounding and bias (cf. chapter ▸Confounding and Interaction of this handbook), the determination of whether an association is causal will also depend on *temporality*, the *strength of the association*, the presence or absence of a *dose-response relationship*, *consistency* with prior literature, and *biological plausibility* (Gordis 2000; Hill 1965; US Department of Health and Human Services 2004).

If an exposure is believed to cause the disease in question, this exposure must occur before the disease develops. *Temporality* is easier to establish for prospective cohort studies in which exposure information prior to disease outcome is available. For cross-sectional or case-control studies, exposure information is usually collected concurrently with disease information or has to be recreated from historical records of exposure, making the assessment of temporality more difficult.

In general, the larger the magnitude of the association, the more likely it is that the relationship between the exposure and disease is causal. In epidemiological studies, the *strength of the association* is usually measured by the relative risk or odds ratio.

If it can be demonstrated that increasing the dose of an agent is associated with increased occurrence of disease in a well-defined relationship, this provides more evidence for causality. The absence of a *dose-response relationship*, however, does not preclude a causal relationship since it is possible that no disease develops until a certain exposure level is reached, after which disease can occur ("threshold effect").

Consistent replication of a finding in different study populations provides further evidence for a causal relationship. However, it is possible that an association only occurs in certain population subgroups, in which case it might be seen in some populations but not others.

Before concluding that an association is causal, it is important to consider *biological plausibility*. While it is possible that epidemiological studies can detect associations which are not yet understood on a biological level, attempting to understand how the exposure might cause the disease in question is nonetheless worthwhile.

Once the results of a study have been finalized, the investigators should consider how they plan to communicate the results and to whom. Groups that should be informed, in general, are the study personnel, study participants, and scientific community. If the results of a study warrant immediate action, health care providers and policy makers should also be alerted. While it is important that the media is informed of the results of studies that have relevance to the general public, it is generally prudent to wait until the study is published in a peer-reviewed journal since the process of critical review of a study allows for the identification and correction of key flaws.

A typical study report consists of the following sections: introduction, methods, results, and discussion (Table 14.6).

Regardless of the audience for the report, results should always be placed in context of the uncertainties and limitations associated with the findings. Describing results in terms of adjectives such as "definitive" or "conclusive" should be avoided. Too often, associations which receive much publicity to begin with have to be rescinded in light of further research.

Concise, simple language aids clarity of presentation. For written reports, adequately labeled tables and figures should be used to summarize information when possible. Information presented in tables should not be merely repeated in the text without additional interpretation.

It is important that results of well-designed studies are reported regardless of whether findings are negative or positive. The tendency for positive findings to be highlighted, both in terms of submission and final publication, biases the perception

Table 14.6 Guidelines for preparation of a study report (Adapted from Szklo and Nieto (2000))

Introduction	
Review study rationale	Describe importance of problem Biological plausibility How does this study add to existing literature?
State hypotheses	Specify interactions of a priori interest
Methods	
Describe study population	Methods of recruitment Inclusion and exclusion criteria
Describe data collection	Include accuracy and reliability of procedures, and quality control measures
State criteria for identification of confounders	
Describe statistical methods	Justify categorization of study variables State assumptions of selected model
Results	
Describe rates of participation or response	
Provide descriptive data	Frequency distributions, means, unadjusted differences Stratify by variables of interest, e.g., age, sex Quality control measures
Present results of model	Use most parsimonious model Additive and multiplicative interactions, if present
Tables and figures	Should be self-explanatory Use informative labels and units
Discussion	
Review main study results Describe strengths and limitations of study	Compare and contrast with published literature
Assess bias and confounding	How much would study results be affected by bias/confounding?
Address uncertainty	How precise are the study estimates, given misclassification?
Clinical, public health policy implications	If strength and impact of study results warrants
Future directions	How to improve on study, build on findings

of the true association between exposure and outcome. This is especially problematic in the context of meta-analyses (cf. chapter ►Meta-Analysis in Epidemiology of this handbook) which attempt to quantitatively summarize published studies. A bias toward publishing positive findings results in a biased estimation of overall risk (Easterbrook et al. 1991; Egger and Smith 1998; Ioannidis 1998; Thornton and Lee 2000).

Studies with substantive findings on a research question may have implications for policies related to public health. Researchers may appropriately highlight such findings in their reports, often at the conclusion of the discussion, commenting on the extent to which new knowledge has been generated with policy implications. There has been substantial debate among epidemiological researchers as to whether publications should also make policy recommendations (Samet 2000). In general, policy recommendations should not be made in publications providing research findings, particularly within the constraints of the policy expertise of most researchers and the space that can be devoted to such discussion in an article.

14.5.3 Storage and Retrieval of Data

Commitment to an epidemiological study does not end with the publication of the final papers. After the study is completed, sufficient material should be stored to allow future sharing of the data or auditing of the study. In fact, for some research funded by the US National Institutes of Health, submission of a fully documented dataset is required, although there are complicated issues related to assuring the privacy of all participants. An index of all stored study materials should be created, along with a description of where they can be located. Materials that could be archived include source data and specimens, laboratory or research notebooks, and the study protocol. Also included should be the final study report, computer data files, copies of computer programs and statistical procedures that were used in analysis, and any printouts of analyses that formed the basis of results included in the final report (Freedland and Carney 1992; The Chemical Manufacturers Association's Epidemiology Task Force 1991; US Environmental Protection Agency (EPA) 1989). If applicable, study forms and related forms should be destroyed in accordance with local statutes and medical records. In order to ensure safety and confidentiality of study materials, storage should be in a physically secure place with limited access.

Periodic checking of stored material is recommended to ensure that necessary updates have been made and to avoid unnecessary clutter. Original records can be transferred to microfilm for storage purposes to conserve space. If microfilm is used, the original records should be retained until the microfilm is checked for proper identification and legibility. For very large studies, electronic storage of study data might make sense given space and cost limitations.

14.6 Conclusions

The field of epidemiology has been growing rapidly, with a vast number of epidemiological studies published every year. A search for "epidemiology" in the PUBMED database yielded 2,028 references in 1964. A similar search for the year 2009 yielded 84,592 references (Fig. 14.3). The results of many of these studies, however, are inconsistent. These inconsistencies are sometimes due to chance, but

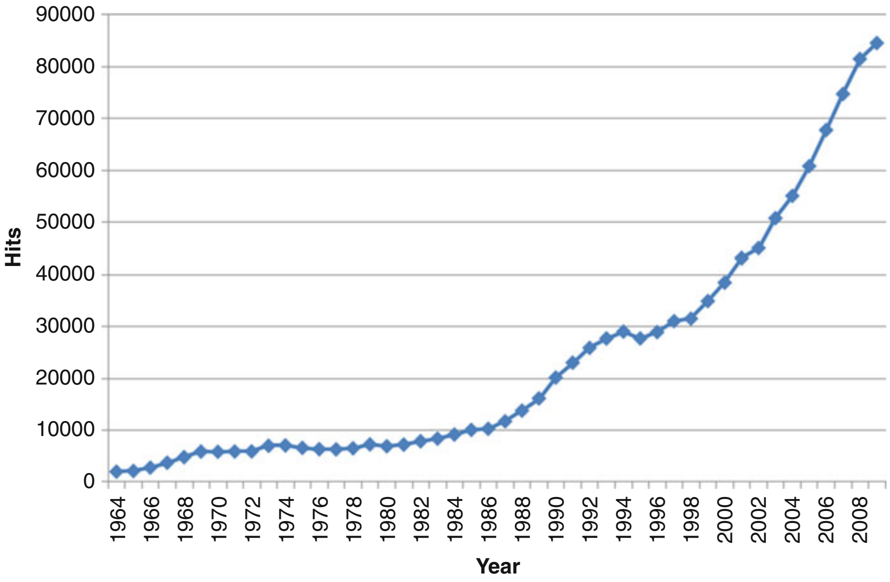


Fig. 14.3 Number of references to "epidemiology" in the PubMed database, 1964–2009

more often can be ascribed to the variable quality of studies with respect to design, conduct, analysis, or dissemination.

As a consequence of the inconsistent results reported by epidemiological studies, many consumers of epidemiological research including clinicians, policy makers, and the general public are dismissive of new findings. The importance of a widespread effort to follow good epidemiological practice and implement rigorous quality assurance and quality control procedures cannot be overstated.

Even as this chapter is being written, the methods of data collection, processing, and storage are changing rapidly as technological innovations emerge. However, the basic principles of good epidemiological practice, data quality assurance, and control will not change. The increasing use of e-mail, PDA, or web-based questionnaires may reduce data error due to data transfer from paper to electronic files, for example, but errors due to poor questionnaire design or data entry (to name just a few sources of error) will still exist. Similarly, electronic processing and storage of data might be helpful in identifying unusual values, but study investigators will still need to review, interpret, and correct these errors.

In this chapter, we have reviewed quality assurance and quality control activities pertinent to the planning, conducting, and reporting of a study. The mental exercise of "optimizing" the dials on the datascope can be useful while conducting epidemiological studies and when considering the results of already published studies. As high-quality research becomes the norm, the field of epidemiology will gain more respect among fellow scientists, policy makers, and the public.

References

Aanensen DM, Huntley DM, Feil EJ, al-Own F, Spratt BG (2009) EpiCollect: linking smartphones to web applications for epidemiology, ecology and community data collection. PLoS One 4:e6968

Agresti A (1990) Categorical data analysis. Wiley, Hoboken

Altman D, Bland J (1983) Measurement in medicine: the analysis of method comparison studies. Statistician 32:307–317

American Association for Public Opinion Research (2000) Standard definitions. Final dispositions of case codes and outcome rates for surveys. American Association for Public Opinion Research, Ann Arbor

Arts DG, De Keizer NF, Scheffer GJ (2002) Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J Am Med Inform Assoc 9:600–611

Asch DA, Jedrziewski MK, Christakis NA (1997) Response rates to mail surveys published in medical journals. J Clin Epidemiol 50:1129–1136

Asch DA, Christakis NA, Ubel PA (1998) Conducting physician mail surveys on a limited budget. A randomized trial comparing $2 bill versus $5 bill incentives. Med Care 36:95–99

Ascherio A, Stampfer MJ, Colditz GA, Rimm EB, Litin L, Willett WC (1992) Correlations of vitamin A and E intakes with the plasma concentrations of carotenoids and tocopherols among American men and women. J Nutr 122:1792–1801

Bachmann J (2007) Will the circle be unbroken: a history of the U.S. National Ambient Air quality standards. J Air Waste Manag Assoc 57:652–697

Baer A, Saroiu S, Koutsky LA (2002) Obtaining sensitive data through the web: an example of design and methods. Epidemiology 13:640–645

Barnett V, Lewis T (1994) Outliers in statistical data. Wiley, Hoboken

Berkanovic E (1980) The effect of inadequate language translation on Hispanics' responses to health surveys. Am J Public Health 70:1273–1276

Bisgard KM, Folsom AR, Hong CP, Sellers TA (1994) Mortality and cancer rates in nonrespondents to a prospective study of older women: 5-year follow-up. Am J Epidemiol 139:990–1000

Blackmore CC, Richardson ML, Linnau KF, Schwed AM, Lomoschitz FM, Escobedo EM, Hunter JC, Jurkovich GJ, Cummings P (2003) Web-based image review and data acquisition for multiinstitutional research. AJR Am J Roentgenol 180:1243–1246

Blomgren KJ, Sundström A, Steineck G, Wiholm BE (2006) Interviewer variability – quality aspects in a case-control study. Eur J Epidemiol 21:267–277

Bray F, Parkin DM (2009) Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. Eur J Cancer 45:747–755

Brenner H, Gefeller O (1997) Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat Med 16:981–991

Breslow N, Day N (1980) Statistical methods in cancer research. Volume I – the analysis of case-control studies. International Agency for Research on Cancer, Lyon

Breslow N, Day N (1987) Statistical methods in cancer research. Volume II – the design and analysis of cohort studies. International Agency for Research on Cancer, Lyon

Bryant AH, Reinert A (2001) Epidemiology in the legal arena and the search for truth. Am J Epidemiol 154(Suppl 12):S27–S35

Canner PL, Krol WF, Forman SA (1983) The coronary drug project. External quality control programs. Control Clin Trials 4:441–466

Canner PL, Borhani NO, Oberman A, Cutler J, Prineas RJ, Langford H, Hooper FJ (1991) The hypertension prevention trial: assessment of the quality of blood pressure measurements. Am J Epidemiol 134:379–392

Centers for Disease Control and Prevention (CDC) (1999) BRFSS summary quality control report. Centers for Disease Control and Prevention, Atlanta

Cherrie J, Schneider T (1998) Validation of a new method for structured subjective assessment of past concentrations. Ann Occup Hyg 43:235–245

Cherrie J, Krantz S, Schneider T, Ohberg I, Kamstrup O, Linander W (1987) An experimental simulation of an early rock woollag wool production process. Ann Occup Hyg 31:583–593

Choi BC, Pak AW, Purdham JT (1990) Effects of mailing strategies on response rate, response time, and cost in a questionnaire study among nurses. Epidemiology 1:72–74

Christiansen DH, Hosking JD, Dannenberg AL, Williams OD (1990) Computer-assisted data collection in multicenter epidemiologic research. The atherosclerosis risk in communities study. Control Clin Trials 11:101–115

Cicchetti DV, Feinstein AR (1990) High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 43:551–558

Clement DL, De Buyzere ML, De Bacquer DA, de Leeuw PW, Duprez DA, Fagard RH, Gheeraert PJ, Missault LH, Braun JJ, Six RO, Van Der Niepen P, O'Brien E, Office versus Ambulatory Pressure Study Investigators (2003) Prognostic value of ambulatory blood-pressure recordings in patients with treated hypertension. N Engl J Med 348:2407–2415

Clive RE, Ocwieja KM, Kamell L, Hoyler SS, Seiffert JE, Young JL, Henson DE, Winchester DP, Osteen RT, Menck HR, Fremgen A (1995) A national quality improvement effort: cancer registry data. J Surg Oncol 58:155–161

Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70:213–220

Collins RL, Ellickson PL, Hays RD, McCaffrey DF (2000) Effects of incentive size and timing on response rates to a follow-up wave of a longitudinal mailed survey. Eval Rev 24:347–363

Comstock GW, Tockman MS, Helsing KJ, Hennesy KM (1979) Standardized respiratory questionnaires: comparison of the old with the new. Am Rev Respir Dis 119:45–53

Cook RR (1991) Overview of good epidemiologic practices. J Occup Med 33:1216–1220

Cooper GR (1986) The importance of quality control in the multiple risk factor intervention trial. Control Clin Trials 7:3

Cottler LB, Zipp JF, Robins LN, Spitznagel EL (1987) Difficult-to-recruit respondents and their effect on prevalence estimates in an epidemiologic survey. Am J Epidemiol 125:329–339

Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, Pratt M, Ekelund U, Yngve A, Sallis JF, Oja P (2003) International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 35:1381–1395

Crombie IK, Irving JM (1986) An investigation of data entry methods with a personal computer. Comput Biomed Res 19:543–550

Data Quality Act of 2000, Public Law 106–554, 114 STAT. 2763, 21 December 2000, Section 515a

Dawber TR, Meadors GF, Moore FE Jr (1951) Epidemiological approaches to heart disease: the Framingham study. Am J Public Health 41:279–286

Day S, Fayers P, Harvey D (1998) Double data entry: what value, what price? Control Clin Trials 19:15–24

Deyo RA, Diehr P, Patrick DL (1991) Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control Clin Trials 12(Suppl 4):142S–158S

Dillman D (1978) Mail and telephone surveys: the total design method. Wiley, New York

Dischinger P, DuChene AG (1986) Quality control aspects of blood pressure measurements in the multiple risk factor intervention trial. Control Clin Trials 7(Suppl 3):S137–S157

Doody MM, Sigurdson AS, Kampa D, Chimes K, Alexander BH, Ron E, Tarone RE, Linet MS (2003) Randomized trial of financial incentives and delivery methods for improving response to a mailed questionnaire. Am J Epidemiol 157:643–651

Dosemeci M, Rothman N, Yin SN, Li GL, Linet M, Wacholder S, Chow WH, Hayes RB (1997) Validation of benzene exposure assessment. Ann N Y Acad Sci 837:114–121

DuChene AG, Hultgren DH, Neaton JD, Grambsch PV, Broste SK, Aus BM, Rasmussen WL (1986) Forms control and error detection procedures used at the coordinating center of the multiple risk factor intervention trial (MRFIT). Control Clin Trials 7(Suppl 3):S34–S45

Eaker S, Bergström R, Bergström A, Adami HO, Nyren O (1998) Response rate to mailed epidemiologic questionnaires: a population-based randomized trial of variations in design and mailing routines. Am J Epidemiol 147:74–82

Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR (1991) Publication bias in clinical research. Lancet 337:867–872

Edwards S, Slattery ML, Mori M, Berry TD, Caan BJ, Palmer P, Potter JD (1994) Objective system for interviewer performance evaluation for use in epidemiologic studies. Am J Epidemiol 140:1020–1028

Edwards SL, Slattery ML, Ma KN (1998) Measurement errors stemming from nonrespondents present at in-person interviews. Ann Epidemiol 8:272–277

Edwards P, Roberts I, Clarke M, DiGuiseppi C, Pratap S, Wentz R, Kwan I, Cooper R (2007) Methods to increase response rates to postal questionnaires. Cochrane Database Syst Rev(2):MR000008

Egger M, Smith GD (1998) Bias in location and selection of studies. BMJ 316:61–66

Etminan M, Gill S, Fitzgerald M, Samii A (2006) Challenges and opportunities for pharmacoepidemiology in drug-therapy decision making. J Clin Pharmacol 46:6–9

Feinstein AR, Cicchetti DV (1990) High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 43:543–549

Fleiss J (1981) Statistical methods for rates and proportions, 2nd edn. Wiley, New York

Fleming TR (1993) Data monitoring committees and capturing relevant information of high quality. Stat Med 12:565–570

Fowler F, Mangione T (1986) Reducing interviewer effects on health survey data. Center for Survey Research, University of Massachusetts, Boston

Fowler F, Mangione T (1990) Standardized survey interviewing: minimizing interviewer-related error. Sage Publications, Newberry Park

Freedland KE, Carney RM (1992) Data management and accountability in behavioral and biomedical research. Am Psychol 47:640–645

Gassman JJ, Owen WW, Kuntz TE, Martin JP, Amoroso WP (1995) Data quality assurance, monitoring, and reporting. Control Clin Trials 16(Suppl2):S104–S136

Gibson PJ, Koepsell TD, Diehr P, Hale C (1999) Increasing response rates for mailed surveys of Medicaid clients and other low-income populations. Am J Epidemiol 149:1057–1062

Gilbart E, Kreiger N (1998) Improvement in cumulative response rates following implementation of a financial incentive. Am J Epidemiol 148:97–99

Gilbert EH, Lowenstein SR, Koziol-McLain J, Barta DC, Steiner J (1996) Chart reviews in emergency medicine research: where are the methods? Ann Emerg Med 27:305–308

Gissler M, Teperi J, Hemminki E, Meriläinen J (1995) Data quality after restructuring a national medical registry. Scand J Soc Med 23:75–80

Goldberg J, Gelfand HM, Levy PS (1980) Registry evaluation methods: a review and case study. Epidemiol Rev 2:210–220

Goldman LR (2001) Epidemiology in the regulatory arena. Am J Epidemiol 154(Suppl 12): S18–S26

Goodman SN (1999a) Toward evidence-based medical statistics, 1: the *P* value fallacy. Ann Intern Med 130:995–1004

Goodman SN (1999b) Toward evidence-based medical statistics, 2: the bayes factor. Ann Intern Med 130:1005–1013

Gordis L (2000) Epidemiology, 2nd edn. W.B. Saunders, Philadelphia

Greenbaum DS, Bachmann JD, Krewski D, Samet JM, White R, Wyzga RE (2001) Particulate air pollution standards and morbidity and mortality: case study. Am J Epidemiol 154(Suppl 12):S78–S90

Halpern SD, Ubel PA, Berlin JA, Asch DA (2002) Randomized trial of 5 dollars versus 10 dollars monetary incentives, envelope size, and candy to increase physician response rates to mailed questionnaires. Med Care 40:834–839

Hawkins N, Evans J (1989) Subjective estimation of toluene exposures: a calibration study of industrial hygienists. Appl Ind Hyg 4:61–68

Hearst N, Hulley SB (1988) Using secondary data. In: Hulley SB, Cummings SR (eds) Designing clinical research. LWW, Baltimore, pp 53–62

Heilbrun LK, Nomura A, Stemmermann GN (1991) The effects of non-response in a prospective study of cancer: 15-year follow-up. Int J Epidemiol 20:328–338

Hiatt RA (2010) The epicenter of translational science. Am J Epidemiol 172:525–529

Hill AB (1965) The environment and disease: association or causation? Proc R Soc Med 58:295–300

Hilner JE, McDonald A, Van Horn L, Bragg C, Caan B, Slattery ML, Birch R, Smoak CG, Wittes J (1992) Quality control of dietary data collection in the CARDIA study. Control Clin Trials 13:156–169

Hoffman SC, Burke AE, Helzlsouer KJ, Comstock GW (1998) Controlled trial of the effect of length, incentives, and follow-up techniques on response to a mailed questionnaire. Am J Epidemiol 148:1007–1011

Holford TR, Stack C (1995) Study design for epidemiologic studies with measurement error. Stat Methods Med Res 4:339–358

Horbar JD, Leahy KA (1995) An assessment of data quality in the Vermont-Oxford trials network database. Control Clin Trials 16:51–61

Hosking JD, Rochon J (1982) A comparison of techniques for detecting and preventing key-field errors. In: Proceedings of the statistical computing section. American Statistical Association, Washington, DC, pp 82–87

Hosking JD, Newhouse MM, Bagniewska A, Hawkins BS (1995) Data collection and transcription. Control Clin Trials 16(Suppl 2):S66–S103

Hunt JR, White E (1998) Retaining and tracking cohort study members. Epidemiol Rev 20:57–70

International Organization for Standardization (2003) ISO 9000:2000, ISO Technical Committee ISO/TC 176

Ioannidis JP (1998) Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA 279:281–286

James J, Bolstein R (1992) Large monetary incentives and their effect on mail survey response rates. Public Opin Q 56:442–453

John EM, Savitz DA (1994) Effect of a monetary incentive on response to a mail survey. Ann Epidemiol 4:231–235

Johnstone FD, Brown MC, Campbell D, MacGillivray I (1981) Measurement of variables: data quality control. Am J Clin Nutr 34(Suppl 4):804–806

Kaaks R, Ferrari P, Ciampi A, Plummer M, Riboli E (2002) Uses and limitations of statistical accounting for random error correlations, in the validation of dietary questionnaire assessments. Public Health Nutr 5:969–976

Kalantar JS, Talley NJ (1999) The effects of lottery incentive and length of questionnaire on health survey response rates: a randomized study. J Clin Epidemiol 52:1117–1122

Kannel WB (2000) Risk stratification in hypertension: new insights from the Framingham Study. Am J Hypertens 13:S3–S10

Karp DR, Carlin S, Cook-Deegan R, Ford DE, Geller G, Glass DN, Greely H, Guthridge J, Kahn J, Kaslow R, Kraft C, Macqueen K, Malin B, Scheuerman RH, Sugarman J (2008) Ethical and practical issues associated with aggregating databases. PLoS Med 5:e190

Kellerman SE, Herold J (2001) Physician response to surveys. A review of the literature. Am J Prev Med 20:61–67

Kessler RC, Little RJ, Groves RM (1995) Advances in strategies for minimizing and adjusting for survey nonresponse. Epidemiol Rev 17:192–204

Khoury M, Bedrosian S, Gwinn M, Higgins J, Ioannidis J, Little J (2010) Human genome epidemiology: building the evidence for using genetic information to improve health and prevent disease, 2nd edn. Oxford University Press, New York

Kiesler S, Sproull L (1986) Response effects in the electronic survey. Public Opinion Q 50:402–413

Kipen HM, Cody RP, Goldstein BD (1989) Use of longitudinal analysis of peripheral blood counts to validate historical reconstructions of benzene exposure. Environ Health Perspect 82:199–206

Kjelsberg MO, Cutler JA, Dolecek TA (1997) Brief description of the multiple risk factor intervention trial. Am J Clin Nutr 65(Suppl 1):S191–S195

Knatterud GL, Rockhold FW, George SL, Barton FB, Davis CE, Fairweather WR, Honohan T, Mowery R, O'Neill R (1998) Guidelines for quality assurance in multicenter trials: a position paper. Control Clin Trials 19:477–493

Krewski D, Burnett RT, Goldberg MS, Hoover K, Siemiatycki J, Abrahamowicz M, White WH (2000) Reanalysis of the Harvard six cities study and the American cancer society study of particulate air pollution and mortality. Investigators' reports parts I and II. Health Effects Institute, Cambridge

Kromhout H, Oostendorp Y, Heederik D, Boleij JS (1987) Agreement between qualitative exposure estimates and quantitative exposure measurements. Am J Ind Med 12:551–562

Kwok R (2009) Personal technology: phoning in data. Nature 458:959–961

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174

Ling PM, Glantz SA (2002) Using tobacco-industry marketing research to design more effective tobacco-control campaigns. JAMA 287:2983–2989

Little RE, Davis AK (1984) Effectiveness of various methods of contact and reimbursement on response rates of pregnant women to a mail questionnaire. Am J Epidemiol 120:161–163

Maclure M, Schneeweiss S (2001) Causation of bias: the episcope. Epidemiology 12:114–122

Maclure M, Willett WC (1987) Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 126:161–169

Maheux B, Legault C, Lambert J (1989) Increasing response rates in physicians' mail surveys: an experimental study. Am J Public Health 79:638–639

Martinson BC, Lazovich D, Lando HA, Perry CL, McGovern PG, Boyle RG (2000) Effectiveness of monetary incentives for recruiting adolescents to an intervention trial to reduce smoking. Prev Med 31:706–713

Maudsley G, Williams EM (1999) What lessons can be learned for cancer registration quality assurance from data users? Skin cancer as an example. Int J Epidemiol 28:809–815

McMahon SR, Iwamoto M, Massoudi MS, Yusuf HR, Stevenson JM, David F, Chu SY, Pickering LK (2003) Comparison of e-mail, fax, and postal surveys of pediatricians. Pediatrics 111(4 Pt 1): e299–e303

McQuade CE, Kutvirt DM, Brylinski DA, Samet JM (1983) A tracking system for conducting epidemiological case-control studies. Comput Programs Biomed 16:149–153

Meinert C, Tonascia S (1986) Clinical trials: design, conduct and analysis. Oxford University Press, New York

Michaels D (2008) Doubt is their product: how industry's assault on science threatens your health. Oxford University Press, New York

Moorman PG, Newman B, Millikan RC, Tse CK, Sandler DP (1999) Participation rates in a case-control study: the impact of age, race, and race of interviewer. Ann Epidemiol 9:188–195

Mullooly JP (1990) The effects of data entry error: an analysis of partial verification. Comput Biomed Res 23:259–267

Neaton JD, Duchene AG, Svendsen KH, Wentworth D (1990) An examination of the efficiency of some quality assurance methods commonly employed in clinical trials. Stat Med 9: 115–123

O'Connor RJ, Giovino GA, Kozlowski LT, Shiffman S, Hyland A, Bernert JT, Carabello RS, Cummings KM (2006) Changes in nicotine intake and cigarette use over time in two nationally representative cross-sectional samples of smokers. Am J Epidemiol 164:750–759

Olson SH (2001) Reported participation in case-control studies: changes over time. Am J Epidemiol 154:574–581

Olson SH, Voigt LF, Begg CB, Weiss NS (2002) Reporting participation in case-control studies. Epidemiology 13:123–126

Paolo AM, Bonaminio GA, Gibson C, Partridge T, Kallail K (2000) Response rate comparisons of e-mail- and mail-distributed student evaluations. Teach Learn Med 12:81–84

Parkes R, Kreiger N, James B, Johnson KC (2000) Effects on subject response of information brochures and small cash incentives in a mail-based case-control study. Ann Epidemiol 10:117–124

Perneger TV, Etter JF, Rougemont A (1993) Randomized trial of use of a monetary incentive and a reminder card to increase the response rate to a mailed health survey. Am J Epidemiol 138:714–722

Piper BG, Lindsey AM, Dodd MJ, Ferketich S, Paul SM, Weller S (1989) Development of an instrument to measure the subjective dimension of fatigue. In: Funk S, Tournquist E, Champagne M, Copp L, Weise R (eds), Key aspects of comfort: management of pain and nausea. Springer, Philadelphia, pp 199–208

Piper BF, Dibble SL, Dodd MJ, Weiss MC, Slaughter RE, Paul SM (1998) The revised Piper Fatigue Scale: Psychometric evaluation in women with breast cancer. Oncol Nurs Forum 25:677–684

Post W, Kromhout H (1991) Semiquantitative estimates of exposure to methylene chloride and styrene: the influence of quantitative exposure data. Appl Occup Environ Hyg 6: 197–204

Prud'homme GJ, Canner PL, Cutler JA (1989) Quality assurance and monitoring in the hypertension prevention trial. Hypertension prevention trial research group. Control Clin Trials 10(Suppl 3):S84–S94

Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'Connor GT, Rapoport DM, Redline S, Robbins J, Samet JM, Wahl PW (1997) The sleep heart health study: design, rationale, and methods. Sleep 20:1077–1085

Reisch LM, Fosse JS, Beverly K, Yu O, Barlow WE, Harris EL, Rolnick S, Barton MB, Geiger AM, Herrinton LJ, Greene SM, Fletcher SW, Elmore JG (2003) Training, quality assurance, and assessment of medical record abstraction in a multisite study. Am J Epidemiol 157:546–551

Rhodes SD, Bowie DA, Hergenrather KC (2003) Collecting behavioural data using the world wide web: considerations for researchers. J Epidemiol Community Health 57:68–73

Rosner B, Spiegelman D, Willett WC (1992) Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. Am J Epidemiol 136:1400–1413

Rothman KJ, Greenland S (1998) Modern epidemiology, 2nd edn. Lippincott-Raven, Philadelphia

Sacks FM, Handysides GH, Marais GE, Rosner B, Kass EH (1986) Effects of a low-fat diet on plasma lipoprotein levels. Arch Intern Med 146:1573–1577

Samet JM (2000) Epidemiology and policy: the pump handle meets the new millennium. Epidemiol Rev 22:145–154

Samet JM (2009) Data: to share or not to share? Epidemiology 20:172–174

Samet JM, Lee NL (2001) Bridging the gap: perspectives on translating epidemiologic evidence into policy. Am J Epidemiol 154(Suppl 12):S1–S3

Samet JM, Zeger SL, Kelsall JE, Xu J, Kalkstein LS (1997) Particulate air pollution and daily mortality: analyses of the effects of weather and multiple air pollutants (The phase IB report of the particle epidemiology evaluation project). Health Effects Institute, Cambridge

Schweitzer M, Asch DA (1995) Timing payments to subjects of mail surveys: cost-effectiveness and bias. J Clin Epidemiol 48:1325–1329

Shahar E, Folsom AR, Jackson R (1996) The effect of nonresponse on prevalence estimates for a referent population: insights from a population-based cohort study. Atherosclerosis risk in communities (ARIC) study investigators. Ann Epidemiol 6:498–506

Shaw MJ, Beebe TJ, Jensen HL, Adlis SA (2001) The use of monetary incentives in a community survey: impact on response rates, data quality, and cost. Health Serv Res 35:1339–1346

Silver RC, Holman EA, McIntosh DN, Poulin M, Gil-Rivas V (2002) Nationwide longitudinal study of psychological responses to september 11. JAMA 288:1235–1244

Slattery ML, Edwards SL, Caan BJ, Kerber RA, Potter JD (1995) Response rates among control subjects in case-control studies. Ann Epidemiol 5:245–249

Sorensen HT, Sabroe S, Olsen J (1996) A framework for evaluation of secondary data sources for epidemiological research. Int J Epidemiol 25:435–442

Spiegelman D, Schneeweiss S, McDermott A (1997) Measurement error correction for logistic regression models with an “alloyed gold standard”. Am J Epidemiol 145:184–196

Spry VM, Hovell MF, Sallis JG, Hofsteter CR, Elder JP, Molgaard CA (1989) Recruiting survey respondents to mailed surveys: controlled trials of incentives and prompts. Am J Epidemiol 130:166–172

Steeh C (1981) Trends in nonresponse rates 1952–1979. Public Opin Q 45:40–57

Stram DO, Langholz B, Huberman M, Thomas DC (1999) Correcting for exposure measurement error in a reanalysis of lung cancer mortality for the colorado plateau uranium miners cohort. Health Phys 77:265–275

Szklo M, Nieto FJ (2000) Epidemiology: beyond the basics. Aspen Publishers, Gaithersburg

The Chemical Manufacturers Association’s Epidemiology Task Force (1991) Guidelines for good epidemiology practices for occupational and environmental epidemiologic research. J Occup Med 33:1221–1229

Thompson WD (1990) Kappa and attenuation of the odds ratio. Epidemiology 1:357–369

Thompson WD, Walter SD (1988) A reappraisal of the kappa coefficient. J Clin Epidemiol 41:949–958

Thornton A, Lee P (2000) Publication bias in meta-analysis: its causes and consequences. J Clin Epidemiol 53:207–216

Tielemans E, Heederik D, Burdorf A, Vermeulen R, Veulemans H, Kromhout H, Hartog K (1999) Assessment of occupational exposures in a general population: comparison of different methods. Occup Environ Med 56:145–151

Turpin J, Rose R, Larsen B (2003) An adaptable, transportable web-based data acquisition platform for clinical and survey-based research. J Am Osteopath Assoc 103:182–186

US Department of Health, Education, and Welfare (DHEW) (1973) Final report of the tuskegee syphilis study Ad Hoc advisory panel. US Public Health Service, Washington, DC

US Department of Health and Human Services (2004) The health consequences of smoking. A report of the Surgeon General. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta

US Department of Health and Human Services (2009) Application for a Public Health Service Grant. PHS 398, Public Health Service, Bethesda

US Environmental Protection Agency (EPA) (1989) Toxic Substances Control Act (TSCA): good laboratory practice standards. 40 CFR Part 792, 34034–34050

Vantongelen K, Rotmensz N, van der Schueren E (1989) Quality control of validity of data collected in clinical trials. EORTC study group on data management (SGDLM). Eur J Cancer Clin Oncol 25:1241–1247

Vardeman SB, Jobe JM (1999) Statistical quality assurance methods for engineers, Wiley, Hoboken

Wacholder S, McLaughlin JK, Silverman DT, Mandel JS (1992) Selection of controls in case-control studies. I. Principles. Am J Epidemiol 135:1019–1028

Wacholder S, Armstrong B, Hartge P (1993) Validation studies using an alloyed gold standard. Am J Epidemiol 137:1251–1258

Wallace JM, Jr, Bachman JG, O’Malley PM, Johnston LD, Schulenberg JE, Cooper SM (2002) Tobacco, alcohol, and illicit drug use: racial and ethnic differences among U.S. high school seniors, 1976–2000. Public Health Rep 117(Suppl 1):S67–S75

Wechsler D (1989) Wechsler preschool and primary scale of intelligence-revised. Psychological Corporation, San Antonio

White E, Hunt JR, Casso D (1998) Exposure measurement in cohort studies: the challenges of prospective data collection. Epidemiol Rev 20:43–56

White E, Armstrong BK, Saracci R (2008) Principles of exposure measurement in epidemiology: collecting, evaluating, and improving measures of disease risk factors. Oxford University Press, New York

Whitney CW, Lind BK, Wahl PW (1998) Quality assurance and quality control in longitudinal studies. Epidemiol Rev 20:71–80

Willett WC, Stampfer MJ, Underwood BA, Speizer FE, Rosner B, Hennekens CH (1983) Validation of a dietary questionnaire with plasma carotenoid and alpha-tocopherol levels. Am J Clin Nutr 38:631–639

Willett WC, Sampson L, Stampfer MJ, Rosner B, Bain C, Witschi J, Hennekens CH, Speizer FE (1985) Reproducibility and validity of a semiquantitative food frequency questionnaire. Am J Epidemiol 122:51–65

Wright P, Haybittle J (1979a) Design of forms for clinical trials (1). Br Med J 2:529–530

Wright P, Haybittle J (1979b) Design of forms for clinical trials (2). Br Med J 2:590–592

Wright P, Haybittle J (1979c) Design of forms for clinical trials (3). Br Med J 2:650–651

Wyatt J (1995) Acquisition and use of clinical data for audit and research. J Eval Clin Pract 1:15–27

Epidemiological Field Work in Population-Based Studies

15

Arlène Fink

Contents

A. Fink (✉)
Public Health - Health Services/Med-GIM, Palisades, CA, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_11,

15.1 Introduction

Field work in epidemiological studies consists of collecting data in natural and experimental settings to answer research questions or test hypotheses about the origins, distribution, and control of disease in populations. Field data can be collected directly and indirectly. Although direct data collection traditionally includes collecting biological samples such as blood and saliva, epidemiologists also collect data about the health of populations by contacting respondents through telephone, mail, or online. To study a community's use of preventative health services (such as influenza vaccinations), for example, a team of epidemiologists can conduct in-person or telephone interviews or administer written, computer-assisted or online surveys. Indirect data collection includes reviewing written, oral, and visual records of respondents' thoughts and actions and observing them in their natural or experimental environment. To study the extent to which a health care system's medical providers adhere to recommended guidelines for preventative health care, for instance, a team of epidemiologists might review a sample of medical records to identify which preventative services were used and by whom. If the team were interested in understanding why preventative services were (or were not) used, it might review transcripts of audio or videotapes of selected physician and patient encounters.

This chapter focuses on providing practical tips on the spectrum of techniques epidemiologists can use in designing and administering reliable and valid non-biological field measures. Although the chapter focuses on direct data collection, many of the principles apply also to indirect data collection.

15.2 Asking for Information: What are the Characteristics of Straightforward Questions and Responses?

Learning how to ask questions in written and spoken form is essential when collecting field data. A straightforward question asks for information in an unambiguous way and extracts accurate and consistent data. Straightforward questions are purposeful, use correct grammar and syntax, and call for one thought at a time with mutually exclusive questions (Sudman and Bradburn 1982; Fink 2002).

15.2.1 Types of Questions

Purposeful Questions Questions are purposeful when the respondent can readily identify the relationship between the intention of the question and the objectives of the study. If the objectives are to find out about the uses of health services, for instance, and some of the study's questions ask about education or place of birth, an explanation is needed of the relationship between questions and objectives. For instance, the introduction, can say something like: "We plan to compare people with differing backgrounds in their use of health services."

Concrete Questions A concrete question is precise and unambiguous. Adding a dimension of time can help make the question more concrete. For instance, rather than asking: "Has a physician *ever* told you that you have hypertension?" ask, "In the *past 12 months* has a physician told you that you have hypertension?"

Complete Sentences Questions should always be stated as complete sentences. Complete sentences express one entire thought, as in Example 15.1.

Example 15.1. Complete sentences
Poor: Place of birth?
Comment: Place of birth means different things to different people. I might give the city in which I was born, but you might tell the name of the country or hospital.
Better: Name the country in which you were born.

Make sure that experts and a sample of potential respondents review all questions even if you are using already existing and validated instrument. Respondents' reading levels and attention spans may vary across studies.

Open and Closed Questions Questions can take two primary forms. When they require respondents to use their own words, they are called open or open–ended. When the answers or responses are preselected for the respondent, the question is termed closed. Both types of questions have advantages and limitations.

An open question is useful when the intricacies of an issue are still unknown, in getting unanticipated answers, and for describing the world as the respondent sees it – rather than as the questioner does. Also, some respondents prefer to state their views in their own words and may resent the questioner's preselected choices. Sometimes, when left to their own devices, respondents provide quotable material. The disadvantage is that unless the team includes a trained anthropologist or qualitative researcher, responses to open questions are often difficult to interpret and compare.

Some respondents prefer closed questions because they are either unwilling or unable to express themselves. Closed questions are more difficult to write than open ones, however, because the answers or response choices must be known in advance. But the results lend themselves more readily to statistical analysis and interpretation. This feature is particularly important in most epidemiological studies which often rely on relatively large numbers of responses and respondents. Also, because the respondent's expectations are more clearly spelled out in closed questions, the answers have a better chance of being more reliable or consistent over time. Example 15.2 shows a closed question.

Example 15.2. A closed question
How often during the past week were you irritable? Circle one.

	Please circle one
Always or nearly always	1
Sometimes	2
Rarely or never	3

15.2.2 Types of Responses

The choices given to respondents for their answers may take three forms (Fink and Kosecoff 1998; McDowell and Newell 1996; Stewart and Ware 1992). The first is called nominal or categorical. (The two terms are sometimes used interchangeably.) Categorical or nominal choices have no numerical or preferential values. For example, asking respondents if they are male or female is the same as asking them to "name" themselves as belonging to one of two categories: male or female.

The second form of response choice is called ordinal. When respondents are asked to rate or order choices, say, from very positive to very negative, they are given ordinal choices. The third form of response choice results in numerical data such as when a respondent is asked to give his or her height or age at the last birthday.

A hypothetical study finding that uses nominal data results in numbers or percentages, as follows:

> Five hundred respondents were interviewed in a study of preventative health. All were asked to indicate whether or not they perform four health-related activities: (1) exercise at least 30 min most days of the week; (2) eat five or six servings of fruits or vegetables daily; (3) smoke; (4) drink no more than one to two alcoholic drinks daily. Of the 500 respondents, 25% reported that they smoked; only 10% stated that they drank no more than one to two drinks daily. None of the respondents reported exercising at least 30 min per day or eating five or six servings of fruits or vegetables.

Ordinal responses are made to fit on a scale that is ordered from positive (*definitely or probably important*) to negative (*definitely or probably not important*). Ordinal data thus are often characterized by counts of the numbers and percentages of people who select each point on a graded scale.

> Of 500 respondents completing the question, 250 (50.0%) rated each preventative health behavior as definitely or probably important.

Field studies often ask for numerical data as when respondents are asked for their birth date. From the date, you can calculate each respondent's age. Age is considered a numerical and continuous measure, starting with zero and ending with the age of the oldest person in the study. Numerical data lend themselves to many statistical operations. Typical findings might appear as follows: The average age of the respondents was 43 years. The oldest person was 79 years of age and the youngest was 23.

15.3 How are Field Study Measures and Questions Organized?

15.3.1 Length

The length of a measure depends upon what you need to know and how many questions you need to ask to get credible answers (Bourque and Fielder 2003a, b). Another consideration is the respondents. How much time do they have available, and will they pay attention? Relatively young children, for example, may only stay still for a few minutes, so shorter interviews, for example, may be better. You must also consider your resources: Longer measures may be more costly to design, validate and administer.

15.3.2 Question Order

All field measures should be preceded by an introduction, and the first set of questions should be related to the topic described in it (see Box 15.1).

Box 15.1. An introduction to a telephone interview and its first question

Hello. I am calling from the Health Clinic. We are surveying people who use the Health Clinic to find out whether it provides satisfactory services. Your name was selected at random from the Clinic's database. Our questionnaire will take no more than 4 min. You can interrupt me at any time. May I ask you the questions? [IF YES, CONTINUE. IF NO, SAY: THANK YOU AND HANG UP.] [CONTINUE HERE:]

The first question asks you about your overall satisfaction with the Health Clinic. Do you consider it [READ CHOICES]

a. Definitely satisfactory
b. Probably satisfactory
c. Probably not satisfactory
d. Definitely not satisfactory

[DO NOT SAY]

e. No opinion or don't know/wrong answer

Note that the interviewer starts off by saying that questions will be asked about satisfaction with the Health Clinic, and the first question calls for a rating of satisfaction.

In general, questions should proceed from easiest to answer and the most familiar to most difficult and least familiar. In a survey of needs for health services, items can first be asked about the respondent's own needs for services, then their family's, community's, and so on.

Questions of recall should also be organized according to their natural sequence. Do not ask very general questions such as: "When did you first start feeling dizzy?"

Instead, prompt the respondent and ask: “In the past 3 months, how often you felt dizzy? Think about the last time you felt dizzy. Was it in the morning, afternoon, or evening?”

Sometimes the answer to one question will affect the content of another. When this happens, the value of the measure may be diminished (Example 15.3).

Example 15.3. Ordering questions

Which question should come first?

a. How efficient is the nursing staff?

Or

b. Which improvements in nursing do you recommend?

Answer: Question b should come before Question a. If it does not, then the respondent might offer suggestions for the improvement of the nursing staff’s efficiency merely because it has been suggested.

Place a few relatively easy-to-answer questions at the end of the measure. When questionnaires, for instance, are long or difficult, respondents may get tired and answer the last questions carelessly or not answer them at all. You can place demographic questions (age, income, gender, and other background characteristics) at the end because these can be answered quickly.

Avoid many items that look alike. Twenty items, all of which ask the respondent to agree or disagree with statements, may lead to fatigue or boredom, and the respondent may give up. To minimize loss of interest, group the questions and provide transitions that describe the format or topic. For instance, say or print something like: “The next set of questions ask about your use of health services.”

Questions that are relatively sensitive should be placed toward the end. Topics such as grooming habits, religious views, and positions on controversial subjects such as abortion and assisted suicide must be placed far enough along so there is reason to believe the respondent is willing to pay attention, but not so far that he or she is too fatigued to answer properly.

Finally, questions should appear to reasonable people to be in a logical order. Do not switch from one topic to another unless you provide a transitional statement to help the respondent make sense of the order.

Here is a checklist of points to consider in selecting the order for the questions in your field measure:

Checklist to Guide Question-Order

✓ For any given topic, ask relatively objective questions before the subjective ones.
✓ Move from the most familiar topics to the least.
✓ Follow the natural sequence of time.
✓ See to it that all questions are independent.
✓ Avoid many items that look alike.
✓ Ask sensitive questions well after the beginning.
✓ Place questions in a logical order.

15.3.3 Aesthetics and Other Concerns

A measure's appearance is important. A self-administered questionnaire that is hard to read can confuse or irritate respondents who may not answer accurately or at all, reducing the reliability of the responses. A poorly designed interview form with inadequate space for recording answers will reduce the efficiency of even the most skilled interviewers.

15.3.4 Branching Questions or Skip Patterns

What happens when you are concerned with getting answers to questions that you know are only appropriate for part of your group? Suppose you were interviewing older adults in a general practice to learn about their medication use. You know that many of these patients are likely to be taking certain kinds of medications such as antihypertensives, NSAIDs, and aspirin. However, some people will be taking all of these medications each day, while others will be taking none.

If you want to ask about a topic that you know in advance is *not* relevant to everyone in the study, you might design a form such as the one in Example 15.4.

Example 15.4. Skip patterns or branching questions

- Do you take any of the following medications (a list is provided)?
  a. No [GO TO QUESTION 4]
  b. Yes
  [IF YES] How often do you take your usual dose (choices are given such as once a day; only when needed)?

OR

- Do you take any of the following medications?
  a. Yes [COMPLETE SECTION A]
  b. No [GO TO SECTION B]

Skip patterns may be confusing to people and should be avoided in self-administered printed questionnaires. Interviewers must be trained to follow skip patterns to ensure accuracy. Online questionnaires are effective vehicles for branching because you can design the software so that the respondent is automatically guided to the appropriate branch. For instance, if the questionnaire tells the respondent, "If no, go to question 6," the respondent who answers "no" will automatically be sent to question 6.

15.4 What Does It Take to Ensure Proper Administration of Field Instruments?

15.4.1 Self-Administered Questionnaires

Self-administered surveys can be designed for completion online or by "paper-and-pencil." They require a great deal of advance preparation and subsequent monitoring to get a reasonable response rate. These questionnaires are given or sent directly to people for completion. Advance preparation, in the form of careful editing and tryouts, is necessary in helping to produce a clear, readable self-administered questionnaire (Bourque and Fielder 2003a). You should always review the returns. Are you getting the response rate you expected? Are all questions being answered? The following is a checklist for using self-administered questionnaires:

Checklist for Using Self-Administered Questionnaires

✓ Mail respondents a letter or email them in advance telling them the purpose of your study. The letter should inform people to expect a questionnaire, explain the importance of the study and the respondent's role, list study supporters and sources of funding, and describe procedures to ensure confidentiality.

✓ Prepare a short, formal letter to accompany the questionnaire form. If you have already sent an advance letter, this one should be very concise.

✓ Offer to send respondents a summary of the findings so they can see just how the data are used. (If you promise this, allocate resources for it.)

✓ If you ask questions that may be construed as personal – such as gender, age, or income – explain why the questions are necessary.

✓ Keep the questionnaire procedures simple. Provide stamped self-addressed envelopes for written, mailed questionnaires. Make sure no special software is needed for online surveys (e.g., to download graphics). If special software is necessary, set up a system for ensuring that all respondents who are eligible for the survey have access to the software.

✓ Keep questionnaires as short as you can. Ask only questions you are sure you need and do not crowd them together. Give respondents enough room to write or check boxes and be sure each question is set apart from the next.

✓ Consider incentives. This may encourage people to respond. Incentives may range from certificates of appreciation, money, and stamps to pens, fuel, and food.

✓ Be prepared to follow up or send reminders. These should be brief and to the point. For mailed and online surveys, it often helps to send another copy of the questionnaire to non-respondents. Do not forget to budget money and time for these additional mailings.

15.4.2 Interviews

Finding Interviewers Interviewers should fit in as well as possible with respondents. They should avoid flamboyant clothes, haircuts, and so on. Sometimes it is a good idea to select interviewers who are similar to respondents in gender, age, or other demographic characteristics.

It is also important that the interviewers be able to speak clearly and understandably. Unusual speech patterns or accents may provoke unnecessarily favorable or unfavorable reactions. The interviewer's way of talking is of course an extremely important consideration in the telephone interview. The interviewer's attitude toward the study and the respondent will influence the results. If the interviewer does not expect much from the interview and sends this message, the response rate and reliability of responses will probably suffer. To make sure you are getting the most accurate data possible, you should systematically and frequently monitor the interviewers' progress (Bourque and Fielder 2003b).

Training Interviewers The key to a good telephone or face-to-face interview is training (Frey 1989). The overall goal of training should be to produce interviewers who know what is expected of them and how to answer questions and also know where to turn if problems arise unexpectedly in the field.

Whether you are training 2 interviewers or 20, it is important to find a time to meet. The advantage of meetings is that everyone can develop a standard vocabulary and share problems encountered in the field.

Once at the training site, trainees must have enough space to sit and write or perform any other activities you will require of them. If you want them to interview one another as practice for their real task, be sure the room is large enough so that two or more groups can speak without disturbing the others. You may even need several rooms.

Trainees should be taken step by step through their tasks and given an opportunity to ask questions. It is also essential to tell them some of the reasons for their tasks so they can anticipate problems and be prepared to solve them. The most efficient way to make sure the trainees have all the information they need to perform their job is to prepare a manual. Here you can explain what they are to do and when, where, why, and how they are to do it.

Conducting Interviews The following are suggested guidelines for conducting interviews:

- Make a brief introductory statement that will describe who is conducting the interview ("Dr. Mary Doe for Armstrong Memorial Medical Center"), tell why the interview is being conducted ("to find out how satisfied you are with our after-surgery program"), explain why the respondent is being called ("We're asking a random sample of people who were discharged from the hospital in the last 2 months"), and indicate whether or not answers will be kept confidential ("Your name will not be used without your written permission").
- Try to impress the person being interviewed with the importance of the interview and of the answers. People are more likely to cooperate if they appreciate the importance of the subject matter.
- Check the hearing and "literacy" of the respondent. Although it is important to stay on schedule and ask all the questions, a few people may have trouble hearing and understanding some of the questions. If that happens, reappraise the eligibility of the respondent (perhaps an interview is not the best method of obtaining reliable data from this respondent; other methods may be more appropriate). Another option is to speak more clearly and slowly.
- Ask questions as they appear in the interview schedule. It is important to ask everyone the same questions in the same way or the results will not be comparable.

Monitoring Interview Quality To make sure you are getting the most accurate data possible, you should monitor the quality of the interviews. This might mean something as informal as having the interviewer call you once a week, or something as formal as having them submit to you a standardized checklist of activities they perform each day. If possible, you may actually want to go with an interviewer (if it is a face-to-face interview) or spend time with telephone interviewers to make sure that what they are doing is appropriate for the study's purposes. To prevent problems, you might want to take some or all of the following steps:

Tips for Ensuring Quality

- Establish a hot line. This means having someone available to answer any questions that might occur immediately, even at the time of an interview. Consider obtaining a toll-free number.
- Provide written scripts. If interviewers are to introduce themselves or the study, give them a script or set of topics to cover. The script may have to be approved by an Institutional Review Board.
- Make sure you give out extra copies of all supplementary materials. If data collectors are to mail completed interviews back to you, for example, make sure to give them extra forms and envelopes.
- Provide an easy-to-read handout describing the purpose of the interview and the content of the questions.
- Provide a schedule and calendar so that interviewers can keep track of their progress.
- Consider providing the interviewer with visual aids. Visual aids may be extremely important when interviewing people in-person whose ability to speak or read may be limited. The preparation of audiovisual aids for use in an interview is relatively expensive and requires that the interviewers be specially trained in using them.
- Consider the possibility that some interviewers may need to be retrained and make plans to do so.

Computer-Assisted Telephone Interviews or CATI Computer-assisted interviewing is becoming increasingly accepted as a useful field work tool. With CATI, the interviewer reads instructions and questions to the respondent directly from the computer monitor and enters the responses directly into the computer (Bourque and Fielder 2003b). The computer, not the interviewer, controls the progression of the interview questions. No paper copies of the interview are produced, thus eliminating the need to find secure storage place for completed questionnaires.

CATI software programs enable the researcher to enter all telephone numbers and call schedules into the computer. When the interviewer logs on, he or she will be prompted with a list of phone numbers to call, including new scheduled interviews and callbacks. For example, suppose the interviewer calls someone at 8 a.m., but receives no answer. The CATI program can automatically reschedule the call for some other time. CATI programs are also available that enable specially trained interviewers to contact respondents with unique needs. For instance, suppose your study sample consists of people who speak different languages. CATI will allow multi lingual interviewers to log on with certain keywords; the computer will then direct them to their unique set of respondents.

The major advantage of CATI is that once the data are collected they are immediately available for analysis. However, having easy access to data may not always be a blessing. Some researchers may be tempted to analyze the data before the completion of data collection, and the preliminary results may be misleading. A main value of easy access to data, certainly in the early stages of data collection, lies in having the means to check on the characteristics of the respondents and to monitor the quality of the CATI interviewers in obtaining complete data.

Intensive interviewer training is crucial when using CATI in field studies. Interviewers must first learn how to use the CATI software and handle computer problems should they arise during the course of the interview. For instance, what will the interviewer do if the computer "freezes"? Further, the interviewer needs to practice answering questions invariably posed by respondents regarding the study's objectives, methods, human subjects' protections, and incentives. In fact, given the complexity of CATI, training may take up to a week. Thus, at the present time, use of CATI should probably be considered primarily when a study is well funded because it is a relatively expensive and specialized form of data collection.

CATI takes two forms. The first, which is most commonly used, consists of a lab or a facility furnished with banks of telephone calling stations that are equipped with computers linked to a central server. The costs of building the lab are extremely high and include assembling soundproof cubicles and having either a master computer that stores the data from the individual computers or linkage to a server. Additional resources are needed to cover the cost of leasing CATI software and hiring a programmer to install it. Training for this type of CATI is expensive, requiring a great deal of practice. There are also numerous incidental costs including those for headsets, seats and desks, instructional manuals, and service contracts for the hardware.

A second type of computer-assisted telephone interviewing system consists of CATI software programs that are run on laptops. With this type of CATI, the researcher only needs a laptop and access to a telephone connected to the Internet. In time, we can expect that the telephone will be superseded by wireless access to the Internet, making this type of telephone interviewing a desirable method for collecting data in the field and sending them to a central server. Moreover, this second type of CATI is appropriate for studies with a variety of funding levels because it is portable and relatively inexpensive. The portability of laptops, however, raises concerns about patient privacy. Laptops are sometimes shared or stolen, providing easy access to confidential respondent data. In anticipation of these concerns, laptops that are used for CATI should be dedicated to a single study, strict privacy safeguards must be enforced, and interviewers must receive special training to ensure proper CATI implementation and respondent protection. In the USA, patient privacy rules have become increasingly strict (e.g., through the Health Insurance Portability and Accountability Act or HIPAA), with costly penalties for violation.

Other Computerized Interview Methods Telephone interviews may not be appropriate for use in epidemiological studies with large populations who are geographically dispersed. These populations often differ in their linguistic, sociocultural, and ethnic traditions. Computerized interview programs are available that standardize the questions and also provide a mechanism for storing and retrieving data for analysis within and across large groups of people (Slimani et al. 2002).

15.5 How Can You Assure a Reliable and Valid Field Measure?

15.5.1 Pilot Testing

Once a field measure has been assembled, it should be tested to determine the ease with which it can be administered and to estimate the accuracy of the data. Pilot testing includes evaluating the logistics of administration as well as the ease of use of the form itself. The purpose of the pilot test is to answer these questions:

Questions Answered by Pilot Testing

- Will the measure provide the needed information? Are certain words or questions redundant or misleading?
- Are the questions appropriate for the respondents?
- Will information collectors be able to use the forms properly? Can they administer, collect, and report information using any written directions or special coding forms?
- Are the procedures standardized? Can everyone collect information in the same way?
- How consistent or reliable is the information?

15.5.2 Reliability and Validity: The Quality of the Measure

A ruler is considered to be a reliable instrument if it yields the same results every time it is used to measure the same object, assuming the object itself has not changed. A yardstick showing that you are 6 feet 1 inch tall today and 6 feet 1 inch 6 months from today is reliable.

People change over time. You may be more tired, angry, and tense today than you were yesterday. People also change because of their experiences or because they learned something new, but meaningful changes are not subject to random fluctuations. A reliable instrument will provide a consistent measure of important characteristics despite background fluctuations. It reflects the “true” score – one that is free from random errors.

A ruler is considered to be a valid instrument if it provides an accurate measure (free from error) of a person’s height. But even if the ruler says you are 6 ft 1 in. tall today and 6 months from now (meaning it is reliable), it may be incorrect, that is, invalid. This would occur if the ruler were not calibrated accurately, and you are really 5 ft 6 in. tall.

If you develop an instrument that consists of nothing more than asking a hospital administrator how many beds are in a given ward, and you get the same answer on at least two occasions, you would have an instrument that is reliable. But if you claim that the same instrument reflects the quality of medical care, you have a reliable measure of questionable validity. A valid measure is always a reliable one, but a reliable one is not always valid (Bernard 2000; Dawson and Trapp 2001).

15.5.3 Ensuring Quality: Selecting Ready-to-Use Measures

One way to make sure that you have a reliable and valid measure is to use one that someone else has prepared and demonstrated to be reliable and valid through careful testing. This is particularly important to remember if you want to survey attitudes, emotions, health status, quality of life, and health beliefs (Stewart and Ware 1992; McDowell and Newell 1996). These factors, and others like them, are elusive and difficult to measure. To produce a truly satisfactory measure of health, quality of life, and human emotions and preferences thus requires a large-scale and truly scientific experimental study.

15.5.4 Reliability

In reviewing a published field instrument (also, in assessing the quality of a homemade form) you should ask the following questions about four types of reliability: test-retest, equivalence, internal consistency, and interobserver reliability.

Test-Retest Reliability Does the instrument have test-retest reliability? One way to estimate reliability is to determine if someone taking the measure gives the same answers on more than one occasion. Test-retest reliability is computed by administering the measure to the same group on two different occasions and then correlating the scores from one time to the next to obtain a correlation coefficient or r value. Usually, to be considered reliable an instrument should obtain a correlation coefficient of at least 0.70 (Stewart and Ware 1992).

You can calculate test-retest reliability for single questions, subsets, or entire measures. For example, suppose you are studying the use of medications in a sample of older adults. The instrument you are using asks this question, "How many prescription medications do you usually take each day?" In order to assess the consistency of the respondents' answers, you would ask the same question twice: at baseline and then a second time, say 2 to 4 weeks later. Assuming medication-use rates in your sample tend to be stable over short periods of time, any differences in responses to the question can be assumed to reflect measurement error and not changes in the use of medications. To calculate test-retest reliability for an entire measure, you would administer its entire set of questions at two different points in time, score it, and then calculate the correlation coefficient for the two scores.

Equivalence Are alternative forms equivalent? If two different forms of a questionnaire are supposed to measure the same attitude, for example, you should make sure that people are likely to obtain the same score regardless of which one they take. If you want to use Form A of the instrument as a premeasured (before an intervention or treatment), for example, and Form B as a postmeasure (after the intervention or treatment), check the equivalence of the two forms to make sure one is not different from the other.

Equivalence reliability can be computed by giving different forms of the instrument to two or more groups that have been randomly selected. The forms are created either by using differently worded questions to measure the same attributes or by reordering the questions. To test for equivalence, you can administer the different forms (reordered or reworded) at separate time points to the same population, or if the sample is large enough, you can divide it in half and administer each of the two alternate forms to half of the group. In either case, you would first compute mean scores and standard deviations on each of the forms and then correlate the two sets of scores to obtain estimates of equivalence. Equivalence reliability coefficients should be at least 0.70.

Internal Consistency Another measure of reliability is how internally consistent the questions are in measuring the characteristics, attitudes, or qualities that they are supposed to measure. To test for internal consistency, you calculate a statistic called coefficient alpha, or Cronbach's alpha, named after the person who first reported the statistic (Anastasi 1982; Bernard 2000). Coefficient alpha describes how well different items complement each other in their measurement of the same quality or dimension.

Many researchers are not at all concerned with internal consistency because they are not going to be using several items to measure one attitude or characteristic. Instead, they are interested in the responses to each item. Decide if your instrument needs to consider internal consistency (see Example 15.5).

Example 15.5. Internal consistency

Internal consistency is important

A ten-item interview is conducted to find out patients' satisfaction with medical care in hospitals. High scores mean much satisfaction; low scores mean little satisfaction. To what extent do the ten items each measure the same dimension of satisfaction with hospital care?

Internal consistency is not important

A ten-item interview is conducted with patients as part of a study to find out how hospitals can improve. Eight items ask about potential changes in different services such as the type of food that might be served, the availability of doctors, nurses, or other health professionals, and so on. One item asks patients for their age, and one asks about education. Since this interview is concerned with views on improving eight very different services and with providing data on age and education of respondents, each item is independent of the others.

Interobserver Reliability: Kappa Kappa is a statistic used to measure interrater (or intrarater) agreement for nominal measures (Cohen 1960). What is a "high" *kappa*? Some experts have attached the following qualitative terms to *kappa*s:

0.0–0.2 = slight; 0.2–0.4 = fair; 0.4–0.6 = moderate; 0.6–0.8 = substantial, and 0.8–0.10 = almost perfect.

Here are some questions to ask about a published field instrument's validity:

15.5.5 Questions to Ask About a Published Instrument's Validity

- Does the instrument have predictive validity? You can validate an instrument by proving that it predicts an individual's ability to perform a given task or behave in a certain way. For example, a medical school entrance examination has predictive validity if it accurately forecasts performance in medical school. One way of establishing predictive validity is to administer the instrument to all students who want to enter medical school and compare these scores with their performance in school. If the two sets of scores show a high positive or negative correlation, the instrument has predictive validity.
- Does the instrument have concurrent validity? You can validate an instrument by comparing it against a known and accepted measure. To establish the concurrent validity of a new measure of quality of care, you could administer the new instrument and an already established, validated instrument to the same group and compare the scores from both instruments. You can also administer just the new instrument to the respondents and compare their scores on it to experts' judgment of the respondents' attitudes. A high correlation between the new instrument and the criterion measure (the established instrument or expert judgment) means concurrent validity. A concurrent validity study is only valuable if the criterion measure is convincing.
- Does the instrument have content validity? An instrument can be validated by proving that its questions accurately represent the characteristics or attitudes that they are intended to measure. An instrument that is designed to measure health beliefs has content validity, for example, if it contains a reasonable sample of facts, words, ideas, and theories commonly used when discussing or reading about the formation of beliefs about disease or health. Content validity is usually established by consulting the literature and by asking experts and prospective respondents whether the questions represent the knowledge, attitudes, and behaviors you want to measure.
- Does the instrument have construct validity? Construct validity means that the instrument measures what it purports to and not something else. Because of the difficulty of obtaining a true measure of the concepts and ideas that characterize epidemiological studies, construct validity must be established experimentally. One method of doing this is to administer the instrument to people who selected experts say exhibit the behavior associated with the construct. Usually, the experts based their judgments on theories that have empirical support and on clinical experience. If the people chosen by the experts to be exemplars of the behavior also obtain a high score (i.e., a higher score means greater evidence of the behavior), then the instrument is considered to have construct validity. This form of validity is usually established after years of experimentation and experience with the measure.

15.5.6 Suggested Guidelines For Pilot Testing

- Try to anticipate the actual circumstances in which the instrument will be conducted and make plans to handle them. For interviews, this means reproducing the training manual and all forms; for online surveys and mailed questionnaires, you have to produce any cover letters, return envelopes, and so on. Needless to say, this requires planning and time and can be costly.
- You can start by trying out selected portions of the instrument in a very informal fashion. Just the directions on a self-administered questionnaire might be tested first, for example, or the wording of several questions in an interview might be tested. This is sometimes called a "cognitive" pretest.
- Choose respondents for the pilot who are similar to the ones who will eventually complete the measure. They should be approximately the same age, with similar education, and so on.
- Enlist as many people in the trial as seems reasonable without wasting your resources. Probably fewer people will be needed to test a 5-item questionnaire than a 20-item one.
- For reliability, focus on the clarity of the questions and the general format of the instrument. Look for:
  - Failure to answer questions.
  - Giving several answers to the same question.
  - Writing comments in the margins.

Any one of these is a signal that the measure may be unreliable and needs revision. Are the choices in forced-choice questions mutually exclusive? Have you provided all possible alternatives? Is the questionnaire or interview language clear and unbiased? Do the directions and transitions make sense? Have you chosen the proper order for the questions? Is the questionnaire too long or hard to read? Does the interview take too long? For instance, you planned for a 10 min interview, but your pilot version takes 20.

Consider this: In a pilot of a self-administered survey of children's health behaviors, respondents were asked how often they washed their hands after eating. All six children between 8 and 10 years of age answered "always" after being given the choices "always," "never," and "I don't know." The choices were changed to "almost always," "usually," and "almost never." With the new categories, the same six children changed their answers to two "almost always" and four "usually."

15.6 Field Work Language and Culture

If you plan on translating an existing data collection instrument or measure, do not assume that you can automatically reword each question into the new language. Between the original language and the next language often lie cultural gaps. You may need to reword each question.

To avoid confusing people and even insulting them because you misunderstand their language or culture, you should follow a few simple guidelines. These involve enlisting the assistance of people who are fluent in the language (and its dialects) and pilot testing the measure with typical respondents. Suggested guidelines for translation always include the following.

15.6.1 Suggested Guidelines for Translation

- Use fluent speakers to do the first translation. If you can, use native speakers. If you can afford it, find a professional translator. The art of translation is in the subtleties – words and phrases that take years and cultural immersion to learn. If you use fluent speakers, you will minimize the time needed to revise question wording and response choices.
- Tryout the translated measure with three to five native speakers. Ask: What is this question asking you? Can you think of a better way to ask this question?
- Revise the measure with the help of the original translator.
- Translate the measure back into the original language. Use a different translator for this task. Does this "back translated" instrument match the original version? If not, the two translators should work together to make them match.
- Try the resulting measure on a small group (5–10) of target respondents. If the two translators could not agree on wording, let the group decide.
- Revise the measure.
- Pilot test the measure.
- Produce a final version.

15.7 Managing the Data

Data management consists of the methods used to store and organize information so that it can be analyzed. Data management (see also chapter ▶Data Management in Epidemiology of this handbook) starts with an analysis plan and ends with the final analytical operation. The analysis plan describes the hypotheses to be tested or research questions that will be answered. The plan is a guide to the data that will be collected, entered, and subsequently analyzed (Example 15.6).

Example 15.6. A portion of an analysis plan for an interview on health and alcohol use in the elderly
Hypothesis: More men than women will exceed drinking limits.
Variables: gender; drinking limits
Planned Analysis: Chi square to test for differences between numbers of men and women who exceed limits

Modifications to the original analysis plan can be expected, especially in large studies with a great deal of data.

A second data management activity is the creation of a code book. The contents of a code book may vary among researchers. Some researchers include in their definition only a description of the study's variables (such as [DRINK]) and how they are categorized or labeled (such as 1 = 1-2 drinks daily; 2 = 3 or more drinks daily; 9 = no data). Increasingly, many investigators are promulgating the view that code books should include all the information needed to reproduce the study. For example (Example 15.7), the Field Survey, a large California polling survey group, posts information such as below on its web site (http://field.com/fieldpoll/).

Example 15.7. Table of contents for a code book

I. Methods.

  A. Sampling.

    1. Sampling design to include eligibility criteria (e.g., 65 years of age and older; have had at least one drink in the past month.)
    2. Sampling strategies (e.g., stratified random sampling; convenience or opportunistic sampling; etc.)
    3. Sample size and justification.
    4. Recruitment and enrollment.
    5. Sampling statistics to include weight and sampling error calculations.

  B. Human subjects: Informed consent.

  C. Research design or how participants were assigned to groups (if appropriate); number and timing of instrument administration.

II. Data Collection.

  A. A copy of the instrument.

    6. The origins of the questions (e.g., adapted from a published instrument; created for this one.)
    7. Description of how each response is coded (e.g., 1 = yes; 2 = no; 9 = no data.)

  B. Training of data collectors; quality control.

  C. Information on reliability and validity.

III. A Data File Description.

The variable names [DRINK], labels (quantity and frequency) and values and value labels (1 = 1–2 drinks daily; 2 = 3 or more drinks daily; 9 = no data.)

A major problem in data management is how to handle missing data. Say, you mail 100 questionnaires and get 95 back. Is this a 95% response rate? Suppose that upon close examination, you discover that half the respondents did not answer question 5, and that none of the questions was answered by all respondents. With all that missing information, you cannot claim to have a 95% response rate.

What should be done about missing responses? In some situations, you may be able to go back to the respondents and ask them to answer the questions they omitted. In small studies, where the respondents are known, the respondents may be easily recontacted. But collecting information a second time is usually impractical, if not impossible, in most studies. Some studies are anonymous, and you do not even know who the respondents are. In institutional settings, you may have to go back to the Institutional Review Board to get permission to contact the respondents a second time. This may take too much time for your purposes.

Computerized questionnaires can be programmed so that the respondent must answer one question before proceeding to the next. Some respondents may find this approach frustrating, however, and refuse to complete the questionnaire. Although compelling respondents to answer all questions is touted as a major advantage of computer-assisted data collection, some researchers believe that forcing respondents to answer every question is coercive and unethical.

A key management activity is data entry, that is, the process of getting data into the computer. It usually takes three forms. In the first, someone enters data from a coded instrument into a database management program or spreadsheet. The data are then saved in as text or ASCII files, so that they can be exported into a statistical program like SPSS, SAS, or Stata. A second type of data entry involves entering data directly into a statistical program like SPSS, SAS, or Stata. In the third form of data entry, the respondent or interviewer enters responses directly into the computer. Data entry of this type is associated with computer-based measures including CATI and online surveys. The responses are automatically entered into database management systems or statistical programs (usually through special translation software). Programs are also available that will automatically convert one file format into another (say from SAS to Stata).

Database management programs, statistical programs and computer-assisted data collection with automatic data entry can facilitate accuracy by being programmed to allow the entry of only legal codes. For instance, if the codes should be entered as 001–010, then you can write rules so that an entry of 01 or 10 is not permitted. If you try to enter 01 or 10, you will get an error message. With minimum programming, the program can also check each entry to ensure that it is consistent with previously entered data and that skip patterns are respected. That is, the program can make sure that the fields for questions that are to be skipped by some respondents are coded as skips and not as missing data. Designing a computer-assisted protocol requires skill and time. No protocol should be regarded as error free until it has been tested and retested in the field.

Once the data are entered, they need to be cleaned. A clean data set can be used by anyone to get the same results as you do when you run the analysis. Data become “dirty” for a number of reasons including miscoding, incorrect data entry, and missing responses.

To avoid dirty data, make sure that coders or data enterers are experienced, well trained, and supervised. Check variable values against preset maximum and minimum levels, so that if someone enters 50 instead of 5, the maximum, you know there is an error. You can also minimize errors by making sure your coding scheme distinguishes truly missing (no response or no data) from don’t know and not applicable.

Run frequencies on your data as soon as you have about 10% of the responses in. Run them again and again until you are sure that the fieldwork is running smoothly. Frequencies are tabulations of the responses to each question. If your data set is relatively small, you can visually scan the frequencies for errors. For large databases with many records, variables, skip patterns, and open-ended text responses, you may need to do a systematic computerized check. All leading statistical programs provide for cleaning specifications that can be used during data entry and later as a separate cleaning process.

Several other problems may require you to clean up the data. These include having to deal with the complete absence of data because some questionnaires have not been returned, for instance, with missing data from questionnaire that have been returned, and with questionnaires that contain data that are very different from the average respondents.

15.8 What Are Reasonable Resources?

Fieldwork resources are reasonable if they adequately cover the financial costs of and time needed to conduct all activities in the time planned. This includes the costs of, and time for, hiring and training staff, preparing and validating forms, administering the instrument or measure, and analyzing, interpreting, and reporting the findings.

How much does it cost to conduct fieldwork? This question can be answered by obtaining the answers to seven other questions.

1. What are the major tasks?
2. What skills are needed to complete each task?
3. How much time do I have?
4. How much time does each task take?
5. Whom can I hire to perform each task?
6. What are the costs of each task?
7. What additional resources are needed?

Here is a checklist of typical field work tasks:

Field Work Task Checklist

✓ *Prepare the instrument for use in the field.*
  - Identify existing and appropriate instruments.
  - Conduct a literature review.
  - Contact other researchers.
  - Adapt some or all questions from existing instruments.
  - Prepare a new instrument.

✓ *Identify, enroll, and recruit subjects.*
  - Determine eligibility criteria (who should be included and excluded).
  - Determine sample size.
  - Identify sources for identifying respondents (e.g.., existing data bases, patients in a waiting room, patients with appointments).
  - Devise plans for coordinating respondents' willingness to participate and the study's field work needs. For instance, you may have to provide transportation for interviewers or participants.

✓ *Prepare documents for the IRB (Institutional Review Board).*
  - Develop procedures for insuring ethical principles of research. Such principles include respect for people and their autonomy, protecting people from harm and taking active steps to protect them, and balance potential risks with benefit from participation in the study.
  - Devise methods for ensuring the protection of the people who participate in field studies, including the preparation of fliers, recruitment letters, and informed consent forms.
  - Make provisions for protection against research misconduct including exaggerating findings or releasing them without permission.

✓ *Pretest the instrument.*
  - Identify a relatively small sample for the pretest.
  - Conduct a "cognitive" pretest by going over the instrument question by question with each respondent.

✓ *Pilot test the instrument.*
  - Identify the sample for the pilot test.
  - Obtain permission for the pilot test.
  - Analyze the pilot-test data.
  - Revise the instrument to make it final.

✓ *Administer the instrument.*
  - Hire staff.
  - Train staff.
  - Monitor the quality of administration.
  - Retrain staff.
  - Send out mail, supervise the questionnaire, conduct interview.
  - Follow up.

✓ *Manage the data.*
  - Code responses.
  - Prepare code book.
  - Consult with programmer.
  - Train data enterers.
  - Enter the data.
  - Run a preliminary analysis.
  - Clean the data.
  - Prepare a final codebook.

✓ *Analyze the data*.
  - Prepare an analysis plan.
  - Analyze the reliability and validity of the instrument.
  - Analyze the results of the study.

✓ *Report the results*.
  - Write the report.
  - Have the report reviewed.
  - Modify the report based on the reviews.
  - Prepare presentation.
  - Present the report orally.

15.8.1 Who Will Do It, and What Resources are Needed? Personnel, Time, and Money

Fieldwork happens because one or more persons are responsible for completing the required tasks. In a very small study, one or two persons may be in charge of developing field instruments, administering them, analyzing the data, and reporting the results. In larger studies, teams of individuals with differing skills are involved. Sometimes, a study is planned and conducted by the staff with the assistance of consultants who are called in for advice or to complete very specific activities.

First, you need to plan the activities and tasks that need to be completed. Once this is accomplished, you then decide on the skills required for each task. Next, you decide on the specific personnel or job descriptions that are likely to get you as many of the skills you need as efficiently as possible. For example, suppose your study design requires someone with experience in training interviewers and writing questions. You may just happen to know someone who needs a job and has both skills, but if you do not know the right person, knowing the skills needed will help you target your employment search.

The specific resources that are needed for each study will vary according to its size and scope and the number of skills and personnel needed to accomplish each task. Example 15.8 illustrates the types of skills and resources for a "typical" field study.

Example 15.8. Tasks, skills, and resources: an explanation

1. *Prepare the instrument for use in the field*

   If an instrument is to be adapted from an already existing instrument, expertise is needed in conducting literature reviews to find out if any potentially useful instruments are available. Sometimes, a reasonably good instrument is available: Why spend time and money to prepare an instrument if a valid one exists? It helps to have experience in the subject matter being addressed and to know who is working in the field and might either have instruments or questions or know where and how to get them.

   Selecting items or rewording them to fit into a new measure requires special skills. You must be knowledgeable regarding the respondents' reading levels and motivation and have experience writing questions.

   Preparing an entirely new instrument is daunting. A job description for an instrument writer would call for excellent writing skills and knowledge of all topics being assessed.

2. *Prepare materials for the IRB (Institutional Review Board)*

   US researchers cannot perform research with funds from the US government without approval from an Institutional Review Board or IRB – an independent group of people whose job is to evaluate if proposed research conforms to ethical principles (Brett and Grodin 1991, see also chapter ►Ethical Aspects of Epidemiological Research of this handbook). This is, however, not US specific. All high-resource countries have IRBs to ensure compliance with ethical principles. Most low-resource countries also have ethical reviews at the local or national level. The IRB typically requires a written explanation of the study plans (including rationale, purposes, and methods), the field forms, and one or more informed consent forms. The following informed consent form (see Box 15.2) has been approved by an IRB. As you can see, it provides potential respondents with descriptions of the study's purposes, the nature and characteristics of the tasks that will be required, and describes procedures for ensuring confidentiality.

Box 15.2. Sample consent form

The Prostate Cancer Network (PROCANE)

You are asked to take part in three telephone interviews and three self-administered questionnaires on your general health, your quality of life since being diagnosed with prostate cancer, and the quality of healthcare you have received while in the PROCANE Program. XXX MD, MPH is directing the PROCANE research study. Dr. XXX works in the Department of Urology at the University of YYYY. You are being asked to take part in the interviews and questionnaires because you are enrolled in the PROCANE program. You can choose to take part in this study or not. If you volunteer to take part in this study, you may stop taking part in the study at any time. This will have no effect of any kind on the health care you receive through the PROCANE program.

Disclosure Statement

Your health care provider may be an investigator in this research protocol. As an investigator he/she is interested in both your clinical welfare and your responses to the interview questions. Before entering this study or at any time during the study, you may ask for a second opinion about your care from another doctor who is in no way associated with the PROCANE program. You are not under any agreement to take part in any research project offered by your physician.

Reason for the Telephone Interviews and Self-Administered Questionnaires

The interviews and the questionnaires are being done for the following reason: To find out if the PROCANE program is meeting the needs of the patients enrolled in the program. During the telephone interview, a trained member of the PROCANE staff will ask you a series of questions about:

- Your health.
- Your quality of life since being diagnosed with prostate cancer, and
- The quality of the healthcare you have received while in the PROCANE program.

The self-administered questions will cover the same topics. But, you will be able to answer them on your own.

The PROCANE program will use your answers and the answers from other program participants to find out if the program is providing the right services to its participants and to find out if any changes need to be made to the program.

What You Will Be Asked to Do

If you agree to take part in this study, you will be asked to do the following things:

1. Answer three short (20 min) telephone interviews. The telephone interviews will ask you general questions about your health, your quality of life since being diagnosed with prostate cancer, and the quality of healthcare you have received while in the PROCANE program. You will be called to complete an interview when you first enroll in PROCANE, 6 months after your enrollment, and when you leave the PROCANE program. The interviews will be completed at whatever time is best for you.
   Sample questions:
   - How confident are you in your ability to know what questions to ask a doctor?
   - During the PAST 4 WEEKS, how much did pain interfere with your normal work (including both work outside the home and housework)? Would you say not at all, a little bit, moderately, quite a bit, or extremely?
   - How much of the time during the LAST 4 WEEKS have you wished that you could change your mind about the kind of treatment you chose for prostate cancer?

2. Answer three short (20 min) self-administered questionnaires. The self-administered questionnaires will ask you general questions about your health, your quality of life since being diagnosed with prostate cancer, and the quality of healthcare you have received while in the PROCANE program. The self-administered questionnaires will be mailed to you when you first enroll in PROCANE, 6 months after your enrollment, and when you leave the PROCANE program. The self-administered questionnaires can be completed at whatever time is best for you. A pre-paid envelope will be provided to you in which to return each questionnaire.
   Sample questions:
   - Over the PAST 4 WEEKS, how often have you leaked urine?
   - Overall, how big a problem have your bowel habits been for you during the LAST 4 WEEKS?
   - Overall, how would you rate your ability to function sexually during the LAST 4 WEEKS?

3. If you do not understand a question or have a problem with a self-administered questionnaire, you will be asked to call Ms. AAA at the PROCANE office at **1-800-000-000.** She will be able to assist you.

Possible Risks and Discomforts

You may be sensitive about answering questions that ask about your physical and emotional health or your experiences with the PROCANE program. However, you do not have to answer any question with which you are uncomfortable.

Potential Benefits to Subjects and/or to Society

The purpose of the telephone interviews and self-administered questionnaires is to improve the services that PROCANE provides to the men enrolled in the program. Your responses might lead to changes in the program that would improve the services that PROCANE provides.

Payment for Taking Part

No payment will be given to you for completing the telephone interviews or self-administered questionnaires.

Confidentiality

Any information that is collected from you and that can be identified with you will remain confidential. Your identity will not be revealed to anyone outside the research team unless we have your permission or as required by law. You will not be identified in any reports or presentations. Confidentiality will be maintained in the following ways:

- All interviews and questionnaires will be coded with a number that identifies you. Your name will not be on any of these materials.
- A master list of names and code numbers will be kept in a completely separate, confidential, password-protected computer database.
- All copies of the self-administered questionnaires will be kept in a locked file cabinet in a locked research office.
- All telephone interviews will be recorded in a confidential computer database.
- When analysis of the data is conducted, your name will not be associated with your data in any way.
- Only research staff will have access to these files.

Taking Part and Choosing Not to Take Part in Telephone Interviews and Self-Administered Questionnaires

You can choose whether to take part in this study or not. If you decide to take part in the telephone interviews and self-administered questionnaires you may stop taking part at any time. This will have no effect of any kind on the health care you receive through the PROCANE program. The investigator may withdraw you from this research if circumstances arise which warrant doing so.

Identification of Investigators

If you have concerns or questions about this study, please contact XXX, M.D., MPH, by mailing inquires to Box 000, Los Angeles, CA 900000-9990. He can be also reached at 1-800-000-000.

Rights of Participants

You may choose to end your agreement to take part in the telephone interviews and self-administered questionnaires at any time. You may stop taking part without penalty. You are not giving up any legal claims, rights or remedies because you take part in the telephone interviews and self-administered questionnaires. If you have questions about your rights as a research subject, contact the Office for Protection of Research Subjects, 2107 QQQ Building Box 951694, Los Angeles, CA 90095-1694, (310) 999-9999.

I understand the events described above. My questions have been answered to my satisfaction, and I agree to take part in this study. I have been given a copy of this form.

____
Name of Subject (Please Print)

____
Signature of Subject Date

Many IRBs also require detailed explanations of why each question on an instrument was chosen, how the study's participants were selected, and how you plan to ensure privacy for the respondent.

3. *Identify, recruit, and enroll patients*
Participants in field studies can be identified from existing databases (e.g., Medicare data base, physician specialty membership lists, patient appointment logs), and they can be approached in the field (e.g., a clinic's waiting room). To obtain a valid sample, the research must establish eligibility criteria that describe who will be included into and excluded from the study. Effective procedures must be devised to approach potential participants, screen for eligibility, inform eligible respondents about their role in the study, and enlist their cooperation. Often, these procedures need to be tested and retested until they achieve their desired goals. In the US as in most other countries, any contact with research subjects, including letters informing potential participants about a study or "scripts" to screen or enroll participants, must be approved by an IRB.
Recruitment letters (see Box 15.3) are often sent in advance of a study to inform respondents of the study and its purposes. The following is an example of a recruitment letter for a study of staffing and clinical policies regarding labor and delivery in all hospitals in a single US state. The letter was approved by an IRB and sent to the appropriate nurse manager at all hospitals in the state that delivered babies in a given year.

Box 15.3. Recruitment letter

<Letterhead and Logos>

Address
____
____
____

Date ____________

Dear Colleague (use name)

We are writing to invite you to participate in an exciting federally funded study of the range of clinical policies on Labor and Delivery (L&D) units in this state. Your participation includes taking part in a structured interview regarding policies and procedures on your (L&D) unit.

As you are well aware, much interest has been placed on nurse staffing policies in general, but to date most studies have centered on medical and surgical units. This will be the first study that is specifically designed to identify staffing patterns used on Labor and Delivery units, and to associate these staffing patterns with clinical policies and patient outcomes.

Likewise, there have been numerous isolated studies regarding the role of nursing support in labor, the effectiveness of doulas, nurse midwives, and various techniques of labor management (e.g. active management, "walking epidurals," hydrotherapy, etc). To date, however, there has been no systematic attempt to describe what is really happening on L&D units in the "real world." At the completion of our study, we hope to be able to answer the following types of questions:

- What does "active management" mean to you?
- How prevalent is hydrotherapy?
- What types of clinicians are trained in "teaching" hospitals?
- When do physicians need to be "in house"?
- How are L&D units staffing in the current environment of a nursing shortage?
- What strategies are being used to monitor cesarean rates?
- How are staffing and clinical policies related to maternal and neonatal outcome?

If you agree to participate, we will send you an advance copy of the L&D Clinical Policy Survey, and call to arrange a convenient interview time for you. You will be compensated $60.00 for your participation, which will take about 45 min to 1 h. Responses will be collated and serve as the first comprehensive overview of staffing and clinical policies on L&D units in this state. This project has been reviewed and approved by the Medical Center Institutional Review Board, and an information letter has been included for your review.

This project has been funded by the Agency for Health, and has been widely endorsed by representatives of agencies promoting a better understanding of healthcare practices, healthcare quality, and healthcare outcomes including, but not limited to the following:

- **Mary Smith**, Administrative Director of the Association of this State's Nurse Leaders
- **Tom Rodriquez**, CEO Medical Center, Current Board of the Hospital Association, and Past President of the State Hospital Association
- **Robert Johnson MD, MPH**, Center for Disease Control Director Division of Birth Defects and Developmental Disabilities

- **William Roberts, MD**, College of Obstetricians and Gynecologists, Chair District XXX

If you have any questions regarding this project, please feel free to contact us at 310 666-789. We thank you in advance for your participation in this exciting project.

Sincerely,

Yvonne Bree, RN, DrPH
Vice President & Chief Nursing Officer
YYY Medical Center

Mathilde Grun, MD, MPH
Director Maternal Fetal Medicine & Women's Health Services Research
YYYY Medical Center
Department Obstetrics & Gynecology

Recruitment letters tend to be most effective if they follow these suggestions:

1. Write the letter on letterhead. If possible place one or more logos on the stationery. The logos may be of the university or agency at which the study takes place and one or more of its supporters.
2. Personalize the letter, if appropriate. In the USA, "Dear John," is often handwritten over the "Dear Dr. Jones" to express collegiality.
3. Describe the purpose of the study. In this case, examples of study questions are included.
4. Describe the role each participant will play. This letter informs the respondent that joining the study means participation in an interview that may last up to 60 min.
5. Describe the incentives you are prepared to offer the respondent. In this case, the incentive is financial.
6. Inform the respondent about confidentiality. According to the letter, the Medical Center's Institutional Review Board has approved the study. An information sheet is to be included with this letter describing how confidentiality is to be ensured. If in doubt, or you do not have an information sheet, the letter should include a statement about protection of privacy.
7. Describe the source of funding for the study.
8. Give the names of any agencies or organizations that endorse or co-sponsor the study.

Sometimes, recruitment is done by telephone. As seen below (see Box 15.4), recruitment also means collecting data on refusers. Such data are used to determine if the recruitment approach is effective and to provide information on the similarity between persons who agree to participate and those who do not. If differences exist between participants and refusers, the external validity of the study may be compromised.

Box 15.4. Parent and child telephone recruitment script

Hello, my name is [FIRST AND LAST NAME] and I am calling on behalf of the LAUSD/UCLA study, *Finding Solutions Together*. May I speak to [NAME OF PARENT]?

1. [IF SOMEONE OTHER THAN RESPONDENT ASKS WHY YOU ARE CALLING, SAY:]

I'm calling about a research study being conducted by the school district. [CHECK ONE ANSWER]

a. No one by that name is at this number ⟶ ASK Q2
b. R not available ⟶ SKIP TO Q3
c. I am speaking to the parent or the parent comes to the phone ⟶ SKIP TO Q9
d. Refusal ⟶ SKIP TO Q8

2. [CONFIRM YOU HAVE DIALED CORRECTLY. ASK IF THE RESPONDENT WAS EVER AT THIS NUMBER AND IF THEY HAVE A NEW NUMBER FOR THE PERSON YOU ARE TRYING TO REACH. IF YOUR INFORMANT CANNOT GIVE YOU A NEW NUMBER, TRY DIRECTORY ASSISTANCE FOR A NEW LISTING. IF NO NEW NUMBER IS LISTED, NOTE AS NOT LOCATED.]
3. [SAY]: Is there a more convenient time to reach R?
   a. Yes ⟶ CONTINUE
   b. No ⟶ GO TO Q5
4. [SET CALL BACK APPOINTMENT]
   Date: ____________________
   Time: ____________________
   [SAY]: Okay. We will try back at that time. Thank you. [END CALL].
5. [SAY]: Is this the best number to reach R or do you have a better number for him/her?
   a. Yes ⟶ CONTINUE WITH Q7
   b. No ⟶ GO TO Q6
6. [RECORD NEW NUMBER]:______________________ ⟶ [SAY]: Okay, we will try him/her at this number. Thank you for your help. [END CALL].
7. [SAY]: Okay, thank you. We will try again another time. May I leave you my name and toll free number in case R wants to call me back?
   a. Yes ⟶ PROVIDE NAME AND TOLL FREE NUMBER. END CALL.
   b. No ⟶ THANK THE INFORMANT AND END CALL.
8. [SAY]: Thank you very much for your time. [END CALL. FILL OUT INFORMATION BELOW]:
   Refusal Information
   Who did you speak to? ____
   Reason for refusal?
   ____

Hello, my name is [NAME] (if not already introduced) and I am calling from [NAME OF SCHOOL]. I am calling you today because your child participated in the first part of our study and [NAME OF CHILD] reported on the questionnaire that [HE/SHE] has had difficulties related to stressful experiences that may benefit from our program. I am calling to see if you and your child are interested in participating in a program that can help children learn ways of coping with stressful experiences.

9. [SAY]: Are you interested in hearing more about this study?
   a. Yes ⟶ CONTINUE
   b. No ⟶ END THE CALL⟶ SKIP TO Q13

   We are conducting a study of youth in middle school who have experienced a very stressful event. The goal of the study is to find out ways that young people react to stressful life events, and whether a new program might help them feel better.

   If you and your child volunteer to be in this study, we would ask you to do the following things:

   - Your child would be given the opportunity to attend a group at school for children who have had stressful experiences and who could benefit from learning ways to improve the way that they feel and act. These groups will have five to ten students and a group leader and will meet once a week at school for 10 weeks. The group sessions will be audio taped for research purposes only. There will also be one meeting between your child and the counselor alone, about halfway through the program and four optional parent meetings for parents to learn more about how to help their child at home.
   - Not all children who qualify for this program will be in the program right away. Children will be chosen at random, like a flip of a coin, to start right away in the program or to receive the program in about 3 months. Those children who do not get into the program right away will also be offered other services to help them while they wait to start this school program.
   - In addition to the groups, we will ask that you and your child meet with an interviewer to answer questions about background about your child and family and how your child has been feeling recently. The parent interview will take about 1 h and will be set up at a time and place that is convenient for you. (Fill in description of questions) The child interview will take about 30 min at school (fill in descrip). We will ask you and your child to complete this interview before the program starts, and again at 3 months and at 6 months after starting the program, for a total of three times.
   - We will ask your child's teacher to complete a short checklist about your child's behavior at school before the program, at 3 months, and at 6 months.

10. [SAY]: Are you willing to meet with me to discuss your child's participation in this study in more detail, and if interested, complete the first parent interview?
    a. Yes⟶ GO TO Q11
    b. No⟶ GO TO Q13

11. [SAY]: Would you prefer to meet at the school or at your home?
    [CHECK ONE ANSWER]:

a. School
b. Home [IF HOME, OBTAIN ADDRESS]:
____
____
____

[SAY]: What is a good day and time for [YOU AND/OR YOUR CHILD] to do the interview?
Date:________________
Time:________________
[IF NECESSARY, OTHERWISE, GO TO Q13: SAY]: Do you think you and your child are able to understand and speak English well enough to participate in this program in English?

a. *Yes* ⟶ GO TO Q13
b. *No, cannot speak english well enough to participate* ⟶ SAY: I'm sorry to bother you. Thank you for your interest in participating in this study. Unfortunately, at this time we can only do this program in English. I would be happy to talk to you about other resources where your child can get help. Thank you for your time. [END CALL].

[SAY]: We would like to call you the day before the interview to remind you. Is it OK to call you at this number?
a. *Yes* ⟶ GO TO Q13
b. *No* ⟶ RECORD DIFFERENT NUMBER ⟶ GO TO Q12

12. [RECORD NEW NUMBER]
New number: (____) ______-________ ⟶ GO TO Q16
[SAY]: Thank you for taking the time to speak with me today. [END CALL].

13. [SAY]: Okay, thank you for taking the time to speak with me today. [END CALL. FILL OUT INFORMATION BELOW]:
Refusal Information
Who did you speak to? ________________________________
Reason for refusal? ________________________________
(Questions 14, 15, and 16 omitted to save space.)

4. *Pretest the field instrument*
Pretesting means identifying a relatively small sample of people who are willing to go through each question with you and tell you what it means to them. (This method is called "cognitive pretesting.") Do the participants agree with your interpretation of each question and response? Usually pretests are done using early versions of the study, and so glitches should be expected. You will need to find a secluded place to conduct the pretest, which is almost always an interview. A trained interviewer is needed. Strict rules are needed for recording participants' answers. Experienced personnel are needed to interpret pretest results and translate them into improvements.
5. *Pilot-test the instrument*
Pilot testing means having access to a group of potential respondents that is willing to try out an instrument that may be difficult to understand or complete. Expertise is needed

in analyzing the data from the pilot test, and experience in interpreting respondents' responses is essential. Additional knowledge is needed in how to feasibly incorporate the findings of the pilot test into a more final version of the instrument.

6. *Administer the instrument*

   Face-to-face and telephone or computer-assisted interviews require skilled and trained personnel. Interviewers must be able to elicit the information called for by the interview questionnaire and record or code the answers in the appropriate way. Interviewers must be able to talk to people in a courteous manner and listen carefully. Also, they must talk and listen efficiently. If the interview is to last no longer than 10 min, the interviewer must adhere to that schedule. Interviews become increasingly costly and even unreliable when they exceed their allotted time.

   Among the types of expertise required to put together a mail questionnaire is the ability to prepare a mailing that is user friendly (e.g., includes a self-addressed envelope) and the skill to monitor returns and conduct follow-ups with those not responding. Email surveys also require similar skills. The instrument used to collect data must be user friendly, and you need the skills to keep track of responses and then follow up non-respondents.

   If you plan to conduct online studies, you should consider becoming familiar with commercial software packages that guide survey preparation and analysis. Training in their use may be necessary for projects that do not have a specialist. If the study is being done at a local site (hospital, clinic), then privacy concerns associated with the Web may be especially daunting.

   Expertise is needed in defining the skills and abilities needed to administer the study's field measures and in selecting people who are likely to succeed in getting reliable and valid data. Training is the key. For example, a poorly trained telephone interviewer is likely to get fewer responses than a well-trained interviewer. Because of the importance of training, many large studies use educational experts to assist them in designing instructional materials and programs for training.

   In large and long-term studies, quality must be monitored regularly. Are interviewers continuing to follow instructions? Who is forgetting to return completed interviews at the conclusion of each 2-day session? If deficiencies in the process are noted, then retraining may be necessary.

7. *Manage the data*

   Managing data means programming, coding, and data entry. It also means setting up a database. Programming requires relatively high-level computer skills. Coding can be very complicated, too, especially if response categories are not precoded. Training and computer skills are needed to ensure that data enterers are expert in their tasks. Finally, data cleaning can be a highly skilled task involving decisions regarding what to do about missing data, for example.

8. *Analyze the data*

   Appropriate and justifiable data analysis is dependent on statistical and computer skills. Some studies are very small and require only the computation of frequencies (number and percentages) or averages. Most, however, require comparisons among groups or predictions and explanations of findings. Furthermore, measures of attitudes, values, beliefs, and social and psychological functioning also require knowledge of the statistical methods for ascertaining reliability and validity.

9. *Report the results*

   Writing the report requires communication skills, including the ability to write and present results in tables and figures. Oral presentations require ability to speak in public and to prepare presentations. It helps to have outside reviewers critique the report; time must be spent on the critique and any subsequent revisions. Expenses for reports can mount if many are to be printed and disseminated.

Use the following checklist as a guide in calculating costs and preparing field study budgets.

Costs of Field Work: A Checklist

- ✓ Learn about direct costs. These are all the expenses you will incur *because* of the fieldwork. These include all salaries and benefits, supplies, travel, equipment, and so on.
- ✓ Decide on the number of days (or hours) that constitute a working year. Commonly used numbers in the USA are 230 days (1,840 h) and 260 days (2,080 h). You use these numbers to show the proportion of time or "level of effort" given by each staff member. Obviously these numbers will vary from country to country.
  Example: A person who spends 20% time on the study (assuming 260 days per year) is spending 0.20 × 260, that is, 52 days, or 416 h.
- ✓ Formulate fieldwork tasks or activities in terms of months to complete each.
  Example: Prepare instrument during months 5 and 6.
- ✓ Estimate the number of days (or hours) you need each person to complete each task.
  Example: Jones, 10 days; Smith, 8 days. If required, convert the days into hours and compute an hourly rate (e.g., Jones: 10 days, or 80 h).
- ✓ Learn each person's daily (and hourly) rate.
  Example: Jones, US $320 per day, or US $40 per hour; Smith, US $200 per day, or US $25 per hour.
- ✓ Learn the costs of "benefits" (e.g., vacation, pension, and health) – usually a percentage of the salay.
  Example: Benefits are 25% of Jones's salary. For example, the cost of benefits for 10 days of Jones's time is 10 × 320 per day × 0.25, or US $800.
- ✓ Learn the costs of other expenses that are incurred specifically for *this* study.
  Example: One 2 h focus group with ten participants costs US $650. Each participant gets a US $25 honorarium for a total of US $250; refreshments cost US $50; a focus group expert facilitator costs US $300; the materials costs US $50 for reproduction, notebooks, nametags, and so on.
- ✓ Learn the indirect costs, or the costs that are incurred to keep the study team going. Every individual and institution has indirect costs. Indirect costs are sometimes a prescribed percentage of the total cost of the field work (e.g., 10%).
  Example: All routine costs of doing "business," such as workers' compensation and other insurance; attorney's and license fees; lights, rent, and supplies, such as paper and computer disks.
- If the fieldwork lasts more than 1 year, build in cost-of-living increases.
- Be prepared to justify all costs in writing.
  Example: The purchases include US $200 for 2,000 labels (two per student interviewed) at US $0.10 per label and US $486 for one copy of MIRACLE software for the data management program.

15.9 Conclusions

Fieldwork in epidemiological studies involves collecting information to describe, compare, or explain knowledge, attitudes, and behavior about the health and health care of populations. To assure reliable information, field work depends upon asking straightforward questions. Straightforward questions are purposeful, concrete and expressed as complete questions. Responses may be considered as nominal or categorical, ordinal, and numerical. Open questions allow the respondent to give answers in his or her own words. Coding open responses may be difficult. Closed questions provide the respondent with choices. They are easier to interpret and analyze than open questions but may not provide in-depth information. An instrument's length is dependent upon the resources available to develop and validate a questionnaire. Keep in mind that very long instruments may tire some respondents, thereby reducing the reliability and validity of the results. Questions should be ordered logically and each must be related to the expressed purposes of the study. Relatively simple questions should go first, hardest second. Demographic information is often called for in last place.

Make certain that respondents understand the purposes of the study and each question you plan to ask. If questionnaires are to be completed by mail, include self-addressed envelopes. Try to keep questionnaires as short as possible. For online surveys, avoid the need for the respondent to follow many steps: keep the questionnaire short and easy to use. For all self-administered questionnaires, make sure they are pretested and pilot tested; when possible look at preliminary data to check that all questions are being answered. Interviewing only succeeds with trained interviewers and a method for monitoring the quality of the process. Consider incentives to compensate respondents for their time. Pilot testing is essential to ensure the collection of reliable data. Reliability refers to the consistency with which questions are answered, while validity refers to the accuracy of the answers. Common types of reliability to consider include test-retest and internal consistency. Common types of validity are content, concurrent, predictive, and construct. To improve reliability and validity, check to see that the language and cultural assumptions of the field study are consistent with those associated with the population being studied. Consider using advance letters and incentives to encourage participation and improve response rates. Make certain all measures and the study's logistics are pretested and pilot tested. Regardless of the methods used to collect data in the field, be ever mindful of the need to ensure confidentiality of responses. Field work tends to be costly because of its dependence upon human capital including trained field workers and data managers.

References

Anastasi A (1982) Psychological testing. Macmillan, New York

Bernard HR (2000) Social research methods: qualitative and quantitative approaches. Sage, Thousand Oaks

Bourque LB, Fielder EP (2003a) How to conduct self-administered and mail surveys. Sage, Thousand Oaks/London/New Delhi

Bourque LB, Fielder EP (2003b) How to conduct telephone surveys. Sage, Thousand Oaks/London/New Delhi

Brett A, Grodin M (1991) Ethical aspects of human experimentation in health services research. JAMA 265:1854–1857

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–48

Dawson B, Trapp RG (2001) Basic & clinical biostatistics. Lange Medical Books, McGraw-Hill

Fink A (2002) How to ask survey questions. Sage, Thousand Oaks/London/New Delhi

Fink A, Kosecoff JB (1998) How to conduct surveys: a step-by-step guide. Sage, Thousand Oaks/London

Frey JH (1989) Survey research by telephone. Sage, Newbury Park

McDowell I, Newell C (1996) Measuring health: a guide to rating scales and questionnaires. Oxford University Press, New York

Slimani N, Valsta L, EFCOSUM Group (2002) Perspectives of using the EPIC-SOFT programme in the context of pan-European nutritional monitoring surveys: methodological and practical implications. Eur J Clin Nutr 56(Suppl 2):S63–74

Stewart AL, Ware JE (1992) Measuring functioning and well-being: the medical outcomes study approach. Duke University Press, Durham

Sudman S, Bradburn NM (1982) Asking questions. Jossey-Bass, San Francisco

Exposure Assessment

16

Sylvaine Cordier and Patricia A. Stewart

Contents

S. Cordier (✉)
INSERM U1085 (National Institute of Health and Medical Research), University of Rennes I,
Rennes, France

P.A. Stewart
Stewart Exposure Assessments, LLC, Arlington, VA, USA

Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda,
MD, USA (retired)

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_12,

16.1 Introduction

Accurate exposure assessment is a prerequisite for an efficient study design, more than ever before, because of the increasing challenges that epidemiology has to face to demonstrate low increases in risk, to disentangle mixed potential risk factors in disease causation, and to provide exposure-response relationships for policy makers.

Exposure assessment is the process that leads to establishing a dichotomy between exposed and non-exposed subjects and/or introducing a level of classification between subjects. A prerequisite for any epidemiological study is that there is variability of exposure to the agent of interest within a population and that this variability between subjects (inter-individual variability) will overcome individual variation of exposure (intra-individual variability).

This chapter describes what choices have to be made for a proper exposure assessment depending on the pathological process under study, gives an overview of the different instruments available for this assessment, and highlights some specific difficulties in this process (retrospective assessment, ecological measurement, or multiple exposures). Finally, measurement errors and ways for controlling them are described.

16.2 Definition of Exposure and Exposure Assessment

Exposure can be defined as a contact of an individual with an agent through any medium or environment. An agent can also be thought of as a susceptibility characteristic. The agent is not necessarily considered to be harmful (e.g., exercise or fiber in the diet). *Exposure assessment* aims to identify whether a person is exposed or not (a dichotomous classification) to a particular agent and, if the individual is exposed, to develop a ranking of subjects by exposure level.

16.2.1 Types of Exposure

An exposure may be to a chemical, a biological, a physical, or a societal agent in the external environment (e.g., cadmium, endotoxin, ionizing radiation, and the existence of a support system, respectively). It may be a characteristic of an individual (e.g., weight or physical activity) or a perception of an individual (e.g., lack of control in the workplace). Finally, it may be a biological agent in the body (e.g., herpes virus), a metabolite of an external agent (e.g., 1-hydroxypyrene, a metabolite of polycyclic aromatic hydrocarbons), a substance representing a pathway of action (e.g., DNA-PAH adducts), or the presence of a polymorphism (e.g., NAT wild type). In this chapter, we use the term exposure to apply to all of these, rather than separating external agents from internal agents. The concept of dose is discussed later in this chapter.

16.2.2 True Risk Factor or Surrogate

Ideally, an exposure assessment should focus on the "true" (or causal) risk factor, although it is acknowledged that complete knowledge of contributive causes is not necessary for an effective prevention. In many situations, however, a *surrogate* must be evaluated because the true risk factor has not yet been identified or only a surrogate can be measured. For example, the causal role of inhaled benzo[a]pyrene in the carcinogenicity of cigarette smoking for the lung may never be formally proven because the true risk factor (i.e., the total amount of inhaled benzo[a]pyrene over a period covering many decades) is impossible to measure (Rothman and Greenland 1998). The International Agency for Research on Cancer (IARC) has classified certain work environments as probably carcinogenic to humans, without identifying the specific compound(s) responsible for this health effect (e.g., the process of refining nickel). Thus, although the true risk factor(s) linked to a health effect may not yet be identified or quantified (e.g., nickel refining, tobacco smoking), measurement of a surrogate remains very useful for research and public health purpose. Protective measures and monitoring of exposure can then be implemented. Medical surveillance in the workplace, which usually includes some kind of biological monitoring of compounds known for their toxicity (e.g., urinary cadmium), may be required. A surrogate is useful for identifying factors of variation for the exposure, establishing presumptive causal associations and dose-response relationships, and narrowing the search for the true risk factor(s).

16.2.3 Dose Versus Exposure

The term *exposure* usually refers to contact with an agent in the external environment as indicated above, common nomenclature also may include agents in the body. Measuring an external agent should, but may not, take into account all the exposure sources (e.g., at home, at work, and leisure time), the time spent in each (i.e., activity patterns), and the individual susceptibility to this agent (e.g., due to physical exercise, diet, and physiological and genetic characteristics). These variables will affect the internal *dose* measured in human tissue or fluid. A biological marker of internal dose therefore comes closer to the relevant measure of exposure in some circumstances than an external exposure. This will be discussed more in Sect. 16.3.1. In the rest of the text, the term *exposure* will be used to describe agents that are being estimated for use in exposure- (or dose-) response relationships in an epidemiological study. Dose will be used to describe the level of the true risk factor at the target organ.

16.2.4 Selection of Metric

Once the agent or a scenario to be investigated in the epidemiological study has been selected, the relevant dimensions to quantify this exposure need to be determined. The appropriate quantification of exposure (metric) should reflect the toxic mechanism of action for the agent and disease of interest. The choice of this metric depends on the knowledge about the supposed biological mechanism inducing the health effect. Chronic diseases such as cancer, for example, are thought to be a result of lifetime exposure, so that the exposure metric often studied is *cumulative exposure*, whereas acute diseases such as asthma are thought to be due to recent high exposures, so that the metric often studied is *peak exposures*.

If there is a biological level above which detoxification processes of the organism are impaired (threshold), the *dose rate* (*average*) of an exposure or a peak exposure may be more relevant than cumulative exposure, because exposures below such a threshold would not cause any deleterious effect.

Oftentimes, however, the biological mechanism of the disease process is not known. In such cases, it is useful to explore multiple metrics such as cumulative (lifetime), highest, average (dose rate), highest short-term (peak) exposure, and components of these (e.g., cumulative exposure level or time above a particular exposure level). For example, the induction of carcinogenesis by a mutagenic compound is, theoretically, initiated at any dose, but the mechanism necessitates a long (sometimes several decades) induction period (*latency*). In this case, recent exposure (immediately preceding diagnosis) is not pertinent, and often measurement of past exposure is “lagged,” that is, exposure occurring in years just before diagnosis of the disease is not taken into account. The exposure metric, then, may incorporate a lagged latency. When an adverse effect is expected to occur only above a certain dose (*threshold*), for instance, in acute toxicity, a metric representing a quantitative level above the threshold would be more appropriate than a metric estimating the total exposure.

Often, the total exposure to a given compound received over a particular time period (cumulative exposure) is the relevant parameter in a pathological process. There are, however, several ways to receive the same cumulative exposure: a high intensity for a short period of time or a lower intensity over a longer period. For instance, the history of tobacco smoking is often summarized by a cumulative index (pack-years), that is, the number of years of smoking times the average number of packs of cigarettes smoked every day during the smoking period. This index, or any equivalent based on the product of duration of exposure by an intensity level, does not distinguish between the roles of *duration* of exposure, irrespective of the rate of exposure, and *intensity* of exposure at every instant.

Selection of an exposure metric that does not appropriately describe the pattern of exposure to the agent being investigated as it relates to the disease of interest will result in misclassification and loss of statistical power (see Sect. 16.5).

16.3 Exposure Data

Because exposures can have different natures, the sources of data used in exposure assessments differ. Exposure data can be thought to be of two types: measurement data (*direct*) and *indirect* information (e.g., questionnaire information, diaries, and records of surrogate information).

16.3.1 Measurement Data

Measurement data are generally considered the most accurate type of exposure data because they are objective measures of exposure. Measurement data include measurements of chemical hazards on the skin and chemical or radiation hazards in the food, air, or water in the general environment or in the workplace. They may be measures of quality of life, such as levels of stress. They also include measurements of human health, such as physical activity levels, physiological measurements, such as blood pressure or weight, or measurements of agents in biological tissues, such as drugs or nutrients. They also include measures of internal exposure or effect, such as blood lead levels and DNA adducts, respectively. For more examples of biological markers, please refer to chapter ▸Statistical Methods in Genetic Epidemiology of this handbook.

Measurements may be taken for purposes of an epidemiological study or may be available from existing records. Although individual measurement data are often thought to be the gold standard, they can be subject to substantial biases. Measurements may not represent the intensity of exposure during the relevant time window, for example, current levels of physical activity may not reflect earlier levels of physical activity. The number of measurements on any individual is generally small, and because the variability of some exposures is large (e.g., in air and in water), one or a few measurements may not reflect the metric of interest, such as long-term exposure levels.

In addition, historical measurement data in records may not represent the true exposure level, because the purpose of the data collection was taken for reasons other than to obtain an estimate of the exposure metric of interest to the study investigator. For example, measurements of agents in the workplace often have been taken to evaluate compliance with exposure regulations, and it has been speculated that such data may reflect higher exposures than the true long-term exposure level. Moreover, the analytical method may not have measured the true risk factor (e.g., historical measurements of cholesterol did not distinguish between high- and low-density cholesterol, and many historical measurements of dust in the air did not distinguish respirable dust from inhalable dust).

Biological measurements of exposure (e.g., carbon monoxide in the breath) *or of effect* (e.g., cholinesterase levels in the blood) are generally thought to be the gold standard, because they most closely reflect the dose received by the target organ. (Note that biological measurements can be both exposure data and the outcome, depending on the study design. Here, only biological measurements used as exposure data are discussed.) There are many limitations to this type of measurement, however. The variability of the concentration of an agent in the body is often greater than that seen in the external environment, so that if the number of measurements is limited, a mean of those measurements may not accurately reflect the average exposure. Some biological measurements may not reflect the dose at the target organ. Instead, they may reflect the amount of agent that was not received by the target organ (e.g., if the agent was measured in the urine) or the amount that was metabolized in the body (including by organs other than the target organ). In such cases, it is assumed that the amount measured and the amount in the target organ are highly correlated, but this correlation is likely to vary by agent or by organ and may vary considerably by individual. There are, in addition, no long-term biomarkers for most agents, and current levels may not reflect long-term exposure levels. Moreover, biological measurements reflect the body burden at one point in time. Even if the agent has a long half-life, the measurement may not be an accurate reflection of the total amount received due to metabolism and elimination over time (e.g., McGrail et al. 1995).

Biological measurements are often invasive and costly. For some known risk factors, only invasive techniques are available for biomonitoring, and exposure assessment, therefore, still relies on more traditional instruments. For example, asbestos is a recognized potent carcinogen. One way to evaluate asbestos exposure would be to measure the asbestos in broncho-alveolar lavage specimens. This invasive and expensive technique, however, is not routinely feasible, nor is it appropriate, because it does not reflect past exposure, which is the most relevant for cancer induction. In this example, exposure assessment must rely on indirect methods of measurement such as questionnaires or records.

If the measurement data were taken after the onset of disease (which is very difficult to determine because the onset may not be detectable), the measurements may be an effect of the disease, rather than a precursor. An example of such a measurement is serum levels of androgens and prostate cancer (Hsing 2001).

Because of their cost, biological measurements are used more often in case-control or cross-sectional studies or in a sample of a cohort, rather than for an entire cohort. In spite of these limitations, biological measurements can provide key insights into the toxicological mechanisms of the agent and can be useful in estimating exposure levels if used judiciously. They can be useful in estimating recent or chronic exposure levels that have low variability over time. In addition, they represent concentrations received from all sources of exposure, so that the total amount of exposure received is better estimated. This advantage is especially important when individual work practices, such as hand washing before eating, can affect internal concentrations.

Measurements of the external environment are thought to be a lower gold standard than biological measurements because they do not measure the internal dose received. They too represent only one point in time. This type of measurement often reflects only one source of exposure when several sources may be contributing to a study subject's overall exposure (e.g., pesticide exposures can occur from application at work, in the house and garden, from contamination of the soil from nearby farming operations, and from consumption of pesticide-contaminated food and water). Thus, measurement of only one source may cause other important sources to be missed. Measuring exposures from a single source, therefore, without considering other sources, can result in lower estimates of exposure and an overestimation of disease risk. In addition, external environmental measurements do not provide an estimate of internal dose. There are several advantages of this type of measurement over biological measurements, however. External environmental measurements are non-invasive and less expensive, and the number of agents for which there are analytical methods is larger. The variability of the concentration of an agent in the external environment usually is lower than the intra-individual variability in the body, meaning that when a small number of measurements is available, a small number of environmental measurements on a group of similarly exposed workers are likely to result in a better estimate of the true exposure level than a small number of biological measurements.

Finally, when measurements are taken for the purpose of an epidemiological study, investigators should ensure that the data are collected in a way to reflect the metric being investigated. The sampling strategy should be developed to reflect the goals of the study (e.g., randomly or randomly within strata). Strict quality control methods should be followed. When records of measurements are being used, investigators should review the collection, analytical, and quality control methods to determine the accuracy of the data and how the measurements compare to the metric being assessed in the epidemiological study.

16.3.2 Indirect Exposure Data

The second type of exposure information, indirect data, is derived from questionnaires, diaries, or records identifying measurements of exposure surrogates. Questionnaires may describe measurement data, for example, cigarettes consumed per week or more subjective measures, such as the perception of control at the workplace. Examples of indirect data from diaries or records of surrogates are the amount of milk products consumed or distance of a residence from a hazardous waste site, respectively. As with measurement data, information from questionnaires, diaries, or records may be problematic.

Questionnaires are developed by study investigators to ensure that information is collected in a structured, standardized approach to reduce differential questioning of cases and controls and to ensure that the data are as complete as possible.

The circumstances under which the questionnaire is administered (in person, telephone, mail, at home, or in a hospital) may reflect the level of response. Development and administration of the questionnaire and data entry and cleanup is costly and time-consuming. Computer-assisted personal and telephone interviews (CAPI and CATI, respectively) have substantially reduced data entry and cleanup costs, but their development is more expensive than using a paper copy. They can, however, include logic checks within the questionnaire to catch errors immediately, rather than long after the interview has taken place (cf. chapter ▶Epidemiological Field Work in Population-Based Studies of this handbook). Questionnaires are usually administered by professional interviewers rather than by scientists knowledgeable of the areas being investigated, so that if a respondent asks for clarification or provides a response that is unclear or inappropriate, the interviewer may not be able to respond in a way to increase the quality of the data. Interviewer training and inclusion of probing questions are means to reduce this problem.

In spite of these limitations, oftentimes, questionnaires are the only way to collect information on exposures.

Designing a questionnaire consists of establishing a list of questions in a predefined order, aimed at eliciting the presence of and often the amount of a given exposure. A questionnaire is defined by its content, the time span it covers in a subject's life, and its format and wording. Common sense principles should guide the construction of a questionnaire. Thus, each question and the flow of the questionnaire should be clear and subject to minimal misinterpretation. Administration of the question should not be a substantial burden to the subject, either in regard to the amount of time spent answering the questionnaire, the complexity of the information being collected, or the sensitivity of the questions. One hour is usually considered the maximum amount of time that respondents retain interest, but it may be much less. Aids can be used to help the respondent accurately recall information, such as lists of pesticides, logos, trademarks of products used, and pictures of medication bottles.

The list of questions in the questionnaire should include only those that the respondent can answer and that will ensure an accurate assessment of exposures. As the questions are developed, an analytical strategy also should be developed on how the responses will be used. A minimum set of questions should be asked that ensure maximum efficiency, but a small number of additional questions may be included for cross-checking data. A few "red herring" questions (i.e., questions that are included to determine the accuracy of the responses, such as inserting in a list of real products, a product with a fake name) are often useful to evaluate the responses. More details on conducting interviews can be found in chapter ▸Epidemiological Field Work in Population-Based Studies of this handbook.

The time span of the questionnaire is important. Respondents can more easily report on current exposures than historical ones. Past exposures, however, may be more important than current exposures in the etiology of chronic diseases, but collecting varying information over many years is problematic. Recollection of important life events at the earlier age can improve recall.

The format of the questions will determine the response rate to the question and the accuracy of the response. Open-ended questions (e.g., "What type of exercise did you do when you were in your twenties?") often gather more information than closed-ended questions because respondents can identify important exposures that are not anticipated by the investigator. Open-ended questions, however, require extensive coding, and some information collected is likely to be useless. Furthermore, important exposures may not be recalled. Close-ended questions (e.g., "Did you do any of the following in your twenties: walk? jog? play tennis? etc.") take more time, but the respondent is less likely to forget one of the identified exposures, making the information collected generally more accurate. If, however, the respondent had an important exposure to an agent not on the list, it may not be reported. Open-ended questions may be used in pilot studies to develop more standardized closed formats. Wording should be geared to the educational level of the respondents. In the USA, the reading level of a 14-year-old is generally considered appropriate for general population studies. When developing questionnaires, the investigator should consult one of the many references on questionnaire design (Sudman and Bradburn 1982; Armstrong et al. 1994; cf. chapter ▸Epidemiological Field Work in Population-Based Studies of this handbook).

Screening questions are useful to minimize the time spent on answering inapplicable questions. Screening questions may require a simple yes or no (e.g., “Did you ever take birth control pills?”), or they may be formatted to screen out the lower exposed individuals (e.g., “Did you ever take birth control pills for at least one year?”).

Diaries are another source of exposure information and have been used most frequently for diet and to a lesser extent, physical activity. In a diary, the respondent reports the amount of exposure (e.g., red meat consumption) at some identified frequency (e.g., daily). Diaries are best used when exposure occurs frequently, because if the frequency is too low, the respondent is likely to forget to complete the diary. Time spent recording the information should be minimal (e.g., less than 1 minute), and the time covered by the diary should be short (e.g., 1–2 weeks) to maximize compliance. Diaries should be formatted in a way to ease data entry as much as possible (e.g., check boxes rather than open-ended questions).

Records are often needed for retrospective exposure assessment (see Sect. 16.4.3). Records of surrogate information (including geographic information systems (GIS)) are often used in ecological studies of the environment. Thus, amount of corn grown in various counties may be used to rank individuals with presumed exposure to herbicides. The data in such records may or may not have been accurately collected, but even if the data were accurately collected, the design of the data collection may impact the usefulness of the data in an epidemiological study. For example, the Toxic Release Inventory of the US Environmental Protection Agency collects emissions data from private businesses. These data can be used to identify geographical areas with significant releases of agents into the air, water, and ground. However, there is a minimum amount of contaminant that must be released into the environment before reporting is required. Companies releasing smaller amounts of agents into the environment are not identified. Thus, if there are many small companies of one type in an area, the emissions reported in the database may suggest very low levels that may not, in fact, be low at all. In such cases, there may be no better data available for use in a study, but the protocol and quality control measures for the data collection should be carefully evaluated prior to use of such data, so that the investigator is aware of the strengths and limitations of the data. It may be useful to compare such data to other records systems as well. For example, a study of farmers’ responses on pesticide use found reasonably good agreement with suppliers’ information on pesticides bought by the farmer (Blair and Zahm 1993).

In summary, the choice of a measurement instrument is determined by knowledge of the disease (what is the true risk factor?), the feasibility of the measurement (its invasiveness and the ease of use in the exposure assessment), the cost, and its validity and reproducibility characteristics (see Sect. 16.5).

16.4 The Process of Exposure Assessment

The process of exposure assessment aims at the construction of an individual exposure estimate, from exposure data available, in order to produce a valid and efficient classification of subjects. Exposure data are usually imperfect, however, and there is a need for exposure *assessment* (rather than measurement), in order to approach the relevant dose.

The main steps for building exposure estimates and classification of subjects are described below. The specific problems resulting from the retrospective character of exposure assessment, the use of ecological estimates and the handling of multiple correlated exposures will also be presented, where ecological estimate refers to estimating an exposure level for a group of individuals, rather than for each individual separately.

The process of exposure assessment can be straightforward to relatively complicated, depending on the level of detail and the accuracy of the exposure data (e.g., surrogates of exposure may warrant less-intensive exposure assessment efforts than accurate and detailed exposure information on the true risk factor), the goal of the study (e.g., hypothesis generating or hypothesis testing), and the resources of the investigator.

16.4.1 Creating an Exposure Estimate

Some exposure data need little processing such as information obtained directly from answers to a questionnaire, for example, smoking habits or intake of some kind of nutrients. In other investigations, some type of processing is needed. In the case of diet, for example, food composition tables allow the computation of the amount of nutrients across food groups (e.g., total vitamin A from various fruits, vegetables, meats). These tables take into account the mode of preparation and of preservation of the food. They are usually country-specific and need regular updating for an accurate translation from food groups into nutrients. For more details on assessment of micronutrients, we refer to chapter ▸Nutritional Epidemiology of this handbook.

Similarly, exercise can be measured using an accelerometer that measures movement, so that the total amount of energy expended can be estimated for an individual getting several types of exercise (Ainsworth et al. 1999).

In environmental studies (cf. chapter ▸Environmental Epidemiology of this handbook), the estimation process often is more complicated. These types of studies often make use of recognized pollutant dispersion models using exposure data reported by the subjects as well as exposure data from other records systems. Recently, various approaches relying on geographical-based information systems and land-use regression modeling have been developped to estimate eposure to traffic-related air pollutants (Wu et al. 2011). Estimates of trichloroethylene were developed for a municipal water system in a study of neurobehavioral effects using information on piping, flow input, water demand, and other variables, and a geographic information system (GIS) on the water distribution systems (Reif et al. 2003).

Occupational epidemiology (cf. chapter ▶Occupational Epidemiology of this handbook) also tends to estimate exposure, often retrospective (see Sect. 16.4.3), from multiple pieces of exposure information, but to date, there are no recognized methods. In the past, experts have based their estimates on job titles and industry with little documentation as to how these estimates were derived. Recently, more attention has been paid to identifying *determinants of exposure* (e.g., factors that affect exposure) (Vermeulen et al. 2002). Examples of determinants include the presence of ventilation, the use of protective equipment, and the quantity of the contaminant in the workplace. Models to estimate an exposure score can be developed by simply assigning weights to the values of the determinants. For example, for a study of man-made mineral fiber, type of emission (active, passive), handling of fibers, presence of controls, protective equipment, and other variables were identified as affecting exposures (Cherrie et al. 1996). Variations in these variables across jobs resulted in the assignment of different scores. Use of these determinants in statistical models allows for a more rigorous and transparent estimation process, however, such as for a study of paving workers where measurement data and determinants such as the type of paving (oil, mastic) and the use of tar were used to develop an estimation model for benzo(a)pyrene exposures (Burstyn et al. 2000). Summary estimates used in epidemiological investigations will then combine knowledge on duration, intensity, time since last exposure, in order to create relevant summary measures according to knowledge of the physiological process involved (Kriebel et al. 2007).

16.4.2 Establishing a Level of Classification

In deciding on a classification, a decision must be made as to whether it will be qualitative (yes/no or ever/never), semi-quantitative or ordinal (e.g., low, medium, or high, or scores of say, 1–3, with or without the quantitative levels associated with the categories identified), or quantitative (with units of measurements). This decision is usually based on the quality of the exposure data.

Continuous data (i.e., quantitative) have greater statistical power to find an association than categorical data. Continuous data, however, also provide an impression of higher quality of exposure data than categorical data do, so that if the exposure data are poor, it may be better to describe the exposures categorically.

Oftentimes, investigators believe that *categorical data* are more accurate than continuous data. In one sense, this may be true. It generally is easier to assign a study subject to one of three categories than to estimate a quantitative level. The use of categories, however, does not reduce the error of the exposure assessment because all individuals within the category are assigned the same value. To illustrate this point, when categories are used, either a score is assigned to the category or the median of the range the category represents is used. It would be rare, however, that all individuals within an exposure category actually have the same exposure level. There are likely to be some individuals exposed at the median level of the category who are therefore appropriately assigned. There are also likely to be some individuals on both the low and the high ends of the category who will be assigned the same value as those individuals at the median level. Moreover, the individuals on the edges of adjacent exposure categories (e.g., the individuals on the high end of the low exposure category and the individuals on the low end of the adjacent higher category) are assigned to different exposure categories and therefore to different median values, although they may be very similar in exposure levels. Thus, within any category of exposure, there is variability in exposure levels, and this variability will reduce the ability of the investigator to identify exposure-response relationships.

Another consideration in selecting the level of classification is the underlying assumption of the exposure-response relationship (cf. chapter ▸Analysis of Continuous Covariates and Dose-Effect Analysis of this handbook). Using a continuous measurement of exposure in regression modeling (cf. chapter ▸Regression Methods for Epidemiological Analysis of this handbook) assumes a linear increase of disease risk (or a transformed scale such as logit) for one unit of exposure. Use of categories of exposure, at least as a first approach, will, instead, fit observed values more closely without requiring any hypothesis about the shape of the exposure-response relationship. Categories must be developed, however, keeping in mind the limitations described above.

Grouping Strategies *Exposure groups* are subsets of the population being studied that are viewed as being similarly exposed and therefore assigned the same exposure level. Exposure groups may be defined during questionnaire development, the exposure assessment process, or the analytical stage. When developing questionnaires, exposure groups are defined when responses to the questions are provided in categories. For example, if the possible responses to “At what age did you get your first menstrual period?” are <10, 10–12, 13–14, and ≥15 years of age, these categories result in four exposure groups. In some studies, exposure groups are developed during the exposure assessment process. Thus, in an environmental study, a question may be asked, “How far did you live from the ABC waste site?” The exposure data that will be used in the exposure assessment may be described in three categories, for example, concentrations of an agent within a mile, 2–5 miles, and ≥5 miles. The investigator, then, may develop three exposure groups: one of subjects who report living ≤1 mile, one of subjects living 2–5 miles, and one of subjects living ≥5 miles. Alternatively, the exposure data may be continuous (e.g., concentrations at various distances). In this case, the investigator may leave the question open ended. Alternatively, he/she may prefer to use the same three response categories as indicated above because the investigator may believe that the subjects can more accurately identify the correct category than estimate a continuously measured distance. Finally, during the analytical stage, investigators may decide to group individuals into quartiles or other arbitrary or ad hoc categories. An advantage of this strategy is that categories can be developed using differing cut-points to allow comparisons with other studies.

The definition of exposure groups is important in an epidemiological study because the variability of exposure level within and across groups affects the power to observe an exposure-response relationship (see also Sect. 16.5). There are three types of variability in epidemiological studies. The first is *intra-individual* or day-to-day variability. For example, a subject with a mean alcohol consumption of two glasses a day may have no drinks some days and four drinks other days. The epidemiologist has no control over this variability, but it is important to appreciate that there is variability of most exposures of individuals, which could be important when investigating threshold effects.

Intragroup variability is the variability that occurs within the exposure group. Thus, within an exposure group consuming 2 to 4 drinks/day, there will be some individuals who average two, some who average three, and some who average four drinks/day. *Intergroup variability* is the variability across the groups (e.g., with categories of 0, $\leq$1–2, 3–4, 5–6, and $\geq$6, the range is 0 $\rightarrow$ 6 drinks/day). The more intragroup variability there is compared to the intergroup variability, the more likely that an exposure-response relationship will be missed. The goal, therefore, is to have narrow ranges of exposure levels within the groups (with little to no overlap across other groups due to misreporting) and as wide a range across groups as possible. For example, in a study investigating coal dust and change in lung function (forced expiratory ventilation in one second (FEV_1)), four different exposure groups were evaluated for intragroup and intergroup variability and the effect of variability on the FEV_1. The intragroup variance ranged from 0.18 to 0.35, and the intergroup variance ranged from 0.20 to 0.23 (Heederik and Attfield 2000). The FEV_1 coefficient (in ml per mg/m^3 of coal dust) ranged from -2.0 to -5.9. The exposure group with the lowest intragroup variance (0.18) and the highest intergroup variance (0.23) was associated with the highest loss of FEV_1 per unit of dust exposure (-5.9 ml/mg/m^3 of dust). Intragroup and intergroup variability can be evaluated using analysis of variance techniques (e.g., Burstyn et al. 2000).

16.4.3 Retrospective Exposure Assessment

The challenges of using instruments to measure current (i.e., recent) exposures are compounded when investigating chronic disease. Because historical measurements are often lacking, investigators may collect current measurements and assume that historical levels were similar or extrapolate historical level from the current measurements. Similarly, exposure information is often asked in questionnaires in reference to a single point in time (e.g., 20 years ago or when the subject was at a certain age), which is equivalent to having only one historical measurement. For example, in the area of nutrition, questionnaires used to investigate chronic disease have traditionally collected only information on current diet. Because diets have changed over time, current diet is not necessarily highly correlated to diets of 20 to 30 years ago.

In contrast, in the occupational investigations, however, complete work histories are often collected, which is likely to result in more accurately historical exposure estimates than using only current job. There is a whole body of literature relative to retrospective exposure assessment using job exposure matrices (JEM) or expert assessment from a panel of experts (Benke et al. 2001). A JEM is a cross tabulation of jobs (or job/industry combinations) and agents by time that automatically assigns the same exposure level to all individuals having the same job. Used in association with a subject's complete work history, JEMs or expert evaluation provide an individual probability or intensity of exposure to a given agent.

16.4.4 Ecological Versus Individual Exposure Assessment

Measurement data may not be available on the actual study subject, but rather on individuals thought to be similarly exposed as the individual under study. These types of measurements are called *ecological assessments*. In contrast, assessment of individual exposures takes into account the personal characteristics of the individual. An example of an ecological assessment is assigning the same level of trihalomethanes in a public water supply system to all individuals on that water supply, in spite of the recognition that the concentration of trihalomethanes can vary within a system. Assigning the same exposure level to individuals with different exposure levels will result in misclassification of study subjects because in the same (macro) environment, subjects are likely, in fact, to have different exposure levels. For example, subjects living in an area with a polluted public water supply will be exposed differently to a pollutant in the water depending on whether their water resources come from a public supply or from a private well, the amount of tap water they drink, their use of tap water for cooking, etc.

An ecological evaluation is used when exposure data or resources are limited. Ecological estimates are the rule in areas such as air pollution epidemiology, where individual exposures are often defined by atmospheric measurements at the sampling location nearest to the individual's residence, or more broadly, at the city level. Ecological estimates are also popular in occupational epidemiology, where job exposure matrices have been developed. In these examples, investigators of air pollution or workplace exposures usually do not have measurement data on the individuals or individual-specific parameters such as individual work practices and protective equipment. The ecological evaluation, therefore, assigns the same exposure value to a group of subjects sharing the same (macro) environment.

Ecological evaluations can result in substantial misclassification of exposure levels. In the field of occupation, even among individuals thought by occupational health professionals to have similar exposure levels, the exposure level can be up to three to six times larger or smaller than estimated, as indicated by geometric standard deviations often found (van der Woord et al. 1999). It seems reasonable to assume that similar degrees of misclassification occur among other types of environmental exposures. Extrapolation of measurement data from one individual to another or from a system to an individual therefore must be done with caution.

Ecological measurements are often derived from existing records (air quality monitoring records, occupational measurement surveys) and are much cheaper to obtain and estimate than individual measurements. Using ecological measurement instead of individual measurements makes sense if the contrast of exposure between the groups (e.g., cities or jobs) is greater than variability of exposures among individuals in the same group. Studies based on ecological measurements may also be useful for hypothesis generation.

Individual assessment generally requires a greater assessment effort but is likely to result in less misclassification. Considerations for selecting one approach over the other include the following: time and financial resources, availability of exposure data and its quality and quantity, and the purpose of the study (e.g., hypothesis generating or hypothesis testing and investigation of an exposure-response relationship).

16.4.5 Dealing with Multiple Exposures

In many situations, exposures to various potential risk factors in human populations tend to aggregate for an individual, due to individual behavior. An example is the correlated habits of smoking, alcohol, and coffee drinking among some individuals. Similarly, in the outdoors environment, humans are exposed to mixtures of compounds originating from the same source (e.g., mercury, polychlorinated biphenyls (PCBs), and other organochlorines from eating fish) or from various sources (e.g., carbon monoxide from automobile and truck exhaust).

Epidemiological studies have proved to be informative about many complex mixtures such as cigarette smoke or air pollution. However, identification of the component(s) responsible for the health effects (and their joint effects) observed is still required for a better understanding of disease causation, cost-effective monitoring of the hazard, and an efficient strategy of prevention of disease.

The situation of the mixed exposures cannot be treated as a classical problem of confounding because the exposures are highly correlated. Stratified analysis or multivariate modeling is, in general, inefficient because such analytical approaches do not allow the presence of a high colinearity among different exposures. In addition, the presence of one or several agents representative of mixed exposures or the occurrence of interaction among exposures is not merely a statistical problem. It also requires a strategy that recognizes the different underlying biological hypotheses of the various components of the mixtures. Much of the insight about multiple exposures comes from epidemiology (for instance, tobacco smoke or outdoor air pollution) because toxicological experiments often cannot replicate complex mixtures to which people are exposed across time, and such experiments are usually limited to single components or suitably chosen combinations.

To illustrate the problem of complex mixtures, we describe as an example environmental exposure to PCBs. Similar examples, however, are found in many other areas of study, including diet and occupational exposures. PCBs are a persistent type of industrial compound that includes 209 different chemical members referred to as congeners. The commercial product always is a mixture of correlated congeners, so that studying the toxicity of these compounds is not easy. For example, some PCBs act like dioxins by binding to the aryl hydrocarbon AhR receptor and may result in cancer (Longnecker et al. 1997). Experimental work has shown the highest dioxin-like activity occurs for congeners with no chlorine in the *ortho* position. It has been speculated that neurologic effects of PCBs, on the other hand, may be caused by congeners with chlorine in the *ortho* position.

Samet (1995) has proposed five general strategies for studying such complex mixtures efficiently: (1) treating the mixture as a single agent, (2) selecting an indicator component, (3) creating a summary index, (4) identifying the separate effects of the mixture's individual components, and (5) characterizing the independent and joint effects of the components. We review these strategies with application to the problem of the toxicity of PCBs.

(1) Treating the Mixture as a Single Agent The early studies in Japan and Taiwan that recognized the neurotoxicity of PCBs, and the later studies in Michigan, relied on total PCBs. At that time, congener-specific data were not available (Schantz et al. 2003). The exposure measurements taken in these studies were powerful enough to strongly suggest the neurotoxic potential of PCBs. There is still, however, a debate about discrepancies in health effects among studies in different countries. These discrepancies may be due to different analytical procedures, different patterns of congeners, or different co-exposures to other organochlorines, such as dioxins or furans, which have similar environmental pathways (Longnecker et al. 1997). In summary, treating the mixture of PCBs as a single agent has proved efficient for hazard identification in early work, but exposure misclassification limits the interpretation of the discrepant findings.

(2) Selecting an Indicator Component Several recent large studies have focussed on a small number of congeners present in relatively high concentrations (e.g., PCB 153). The congeners present in high concentrations, however, are not necessarily the most toxic. As a rule, "a single component of a mixture may be an appropriate index of toxicity if the component mirrors the dosimetry and toxicity of other components relevant to the health effect of concern" (Samet 1995).

(3) Creating a Summary Index Creating a summary index implies the attribution of some type of weighting to the individual concentrations of the different components of a mixture. The weight assigned to each congener is defined according to an underlying hypothesis about the biological activity of each component. If one assumes that endocrine disruption is a relevant biological mechanism of toxicity for PCBs, a measurement of the total estrogenic xenobiotic burden in adipose tissue could provide an integrated biomarker of xeno-hormonal activity resulting from exposure to a given mixture of compounds (Soto et al. 1997). Another example of biological activity, the dioxin-like activity of a PCB congener, can be calculated using a toxic equivalency factor (TEF) (Ahlborg et al. 1994), which is assigned relative to the toxicity of the dioxin 2, 3, 7, 8-TCDD. The total toxic equivalency (TEQ) of a mixture of PCBs can then be estimated by summing across all compounds, the product of the concentration, and TEF for each compound. It is likely, however, that the weighting is dependent on the state of knowledge about the relative potency of the different components at the time of calculation and that over time it would be necessary to modify the summary index as more information becomes available.

(4) Separating Effects of the Mixture's Components Creating one summary index does not reflect the heterogeneity of the mixture. There is a trade-off between measuring concentrations of the individual compounds in the mixture (which is usually time-consuming and expensive) and summarizing the mixture of highly correlated congeners. Analyzing concentrations of 38 PCBs congeners from 497 human milk samples from Canada in 1992, Gladen et al. (2003) distinguished three groups of congeners: one group of the congeners, including most of the major congeners, that were highly correlated, meaning that their individual biological effects realistically could not be separated in an epidemiological study; another group of congeners quantifiable in only a small fraction of the population by the assay methods used and therefore an epidemiological analysis would be uninformative; and a third group quantifiable in a reasonable fraction of samples and not correlated with the bulk of major congeners. The authors concluded the components of this last group are worth studying separately and are good candidates for individual determination and inclusion in epidemiological studies.

(5) Characterizing the Independent and Joint Effects of Components Measurements of selected congeners allow the evaluation of health effects related to single or joint exposures. Correlations, however, exist not only between concentrations of PCBs congeners but also with other common organochlorines, metals, and pesticides, and there are strong suspicions of possible interactions among these compounds at the molecular level that affect neurobehavioral function in particular (Carpenter et al. 2002). The strategies presented earlier provide some guidelines for studying these joint effects in epidemiological studies.

Two other points regarding mixtures are appropriate. It should be recognized that while some agents within a mixture may cause a disease, it is possible that other agents in that same mixture reduce the likelihood of the disease by deactivating the active compound. For instance, there is an active discussion around the beneficial impact on birth weight of seafood consumption during pregnancy, which brings high amounts of fatty acids and selenium, relative to the potential toxicity of seafood from contaminants such as mercury (Grandjean et al. 2001). This situation complicates the determination of causality in epidemiological studies. Also, individual characteristics of the study subjects (e.g., polymorphisms) may intensify or reduce the effect of the agent. Currently, our ability to tease out these situations is limited, but investigators should at least recognize that they may be possible.

Multiple exposures can be evaluated using interaction analysis but can also be grouped using hierarchical cluster analysis (e.g., see Hines et al. 1995 for an example). In this study, fabrication workers in a semi-conductor company were exposed to multiple chemicals. Hierarchical analysis allowed the investigators to identify groups of workers exposed to the same pattern of exposures (e.g., various glycol ethers).

16.5 Measurement Errors

All types of exposure assessment in every area of investigation will have some error. Chapter ▸Measurement Error of this handbook describes statistical methods to cope with measurement errors. Appreciation of the types and degree of error allows for a more appropriate interpretation of the study results. Knowing the sources of error can also provide areas for methodological investigation within the study to allow quantification of the error. This, in turn, can allow the investigator to estimate the effect of the error on the epidemiological findings.

16.5.1 Types of Measurement Errors

There are two types of errors that arise from measurements: random and systematic. Random error will result in the measurements being randomly distributed around the mean. Systematic error, or bias, will result in an overall mean that is erroneously high or low compared to the true mean. Both types of error are of concern in exposure assessment, and they are described in terms of precision and validity. *Precision* measures random error and refers to the reproducibility or reliability of the measure. *Validity* measures systematic error and refers to the distance between the exposure measured and the target variable (ideally, the true risk factor, but practically, the surrogate).

A measurement instrument must be *reproducible*. Under ideal conditions, this means that if the instrument is administered under the varying conditions, it should provide the same response within a reasonable level of variation. Generally, however, reproducibility more practically is defined as providing the same response within a reasonable level of variation under the same circumstances. Reproducibility is a necessary condition to accurately evaluate intra-individual and intragroup variability, but somewhat less necessary to accurately evaluate intergroup variability. In addition, to be useful, the measurement instrument must also be *valid* (i.e., it should measure the exposure it is supposed to measure and identify the true quantity present).

Historically, measurement error more often has been associated with categorical assessments than quantitative, probably because quantitative assessments have been limited in the past. Measurement error in either type of assessment will result in *misclassification error* when estimating the exposure levels of study subjects. For example, if a subject was assigned to a high fruit intake category, rather than a medium fruit intake category, the subject is misclassified. Misclassification of confounders can be also a serious problem since it will usually reduce the degree to which confounding can be controlled. For instance, in many studies it is essential to obtain a complete smoking history including detailed periods of smoking or quitting, and quantity smoked during each period, because tobacco smoking is a risk factor and therefore a potential confounder, for many diseases. When studying lifestyle factors associated with smoking, such as alcohol consumption, misclassification of smoking habits will result in incomplete adjustment and residual confounding. In the context of an epidemiological study, misclassification is characterized as non-differential or differential, depending on whether it affects the comparison groups (i.e., the diseased and non-diseased subjects) similarly. *Differential misclassification*, which results from there being a different amount of error for the diseased compared to the non-diseased, can lead to underestimation or overestimation of the association between the exposure and disease. In the latter situation, misclassification can induce spurious statistically significant results. *Non-differential misclassification* of exposure usually will bias estimates of relative risks toward the null. There are examples, however, occurring in extreme conditions, where non-differential misclassification of exposure can produce bias away from the null (Rothman and Greenland 1998). For more discussions about these concepts, see chapter ▸Misclassification of this handbook. Thus, both types of misclassification can result in incorrect conclusions.

16.5.2 Sources of Measurement Errors

Armstrong et al. (1994) classified sources of measurement error in five categories: faulty design of the instrument, errors or omissions in the protocol regarding the use of the instrument, poor execution of the protocol during data collection, limitations due to subject characteristics (e.g., poor memory of past exposures or day-to-day variability in biological characteristics), and errors during data entry and analysis. They have provided an extensive list of circumstances in which these errors may occur, and these sources should be carefully evaluated before attempting to use any type of instrument.

Measurement instruments and analytical methods (such as for an air or biological measurement, blood pressure) generally are designed to be as accurate and reproducible as possible when used under similar conditions, that is, with the same protocol. Two possible sources of systematic differences that can occur are from the measurement/analytical method itself and from the interference of other substances present in the measured environment. Reduction of these errors in the investigation of disease risks can be made by following the manufacturer's/laboratory recommendations, calibrating the instrument under the conditions being measured, using spiked and blank samples, and following other quality control procedures (cf. chapter ▸Quality Control and Good Epidemiological Practice of this handbook). Random error can arise from a lack of technical precision of the instrumentation, variation introduced by the laboratory technicians, and the analytical procedures themselves. This inherent limitation of the instrument and analytical methods, however, explains only part of the variability. Other sources of variation include weather conditions, presence of other exposures, the actual concentration being out of the range of the instrument's measurement range, and the timing of the instrument's response in the relation to a change in concentration. The sources of error need to be identified in order to decrease, or at least, recognize, and quantify the variability.

Questionnaires, because they also can suffer from the two types of misclassification, systematic and random, can be viewed similarly. Systematic differences can result from incorrect phrasing of questions (such that all respondents misinterpret the question similarly) or from inappropriate or misleading response categories. Random sources of misclassification can result from poor phrasing (such that different respondents interpret a question differently) and lack of interest on the part of the respondent. To collect quality data, questionnaires should be standardized, so as to ensure that all study subjects are asked the same questions. Questions must be clearly phrased, without ambiguity, and use terms that are understandable to respondents. Respondents must be able to remember the events being asked about and be able to correctly respond to the questions. Thus, reporting of events that took place many years ago or that require mathematical calculations (e.g., estimating average amount of foods eaten on a seasonal basis) is likely to be subject to more random error than reporting of more recent events or events that do not require calculations (e.g., Bradburn et al. 1987; Subar et al. 1995). Pilot testing of questions should be conducted on a group of individuals with the characteristics of the group who will be receiving the questions because respondents often interpret questions very differently from investigators, even if the questions were carefully developed. Questions should also be tested under the conditions that the questionnaire will be administered (e.g., in the home). Following these procedures should decrease bias and increase precision.

Diaries are prone to both systematic and random errors from the same sources as questionnaires. Records, in contrast, may have systematic and random error similar to measurement data or questionnaire data, depending on the type of record.

Both systematic and random errors may result from limited data. For example, systematic error could result in missing information from asking about sensitive issues, such as the number of sexual partners (Lindzey and Aronson et al. 1985). Subjects may be more inclined to respond with a "don't know" if the number of partners exceeds what they consider to be acceptable. Cases with workplace-induced cancer may be so sick that proxies are used as the respondents. Proxies generally know little about workplaces of the subject. In contrast, many of the control subjects would be able to provide detailed information about the workplace.

Having limited exposure information can result in misclassification of subjects by exposure level. In the environmental area, Brunekreef et al. (1987) illustrated the effect of limited data on misclassification in a study of the relationship between environmental exposure to lead and blood lead levels in children. He found that averaging four measurements of lead on home floors increased the regression coefficient explaining blood lead levels by 69%, compared to the model using a single home floor measurement. Having only one measurement, therefore, would have increased the misclassification of subjects. Generally, non-differential misclassification due to limited data will result in random error.

The problem of limited data also is evident in the use of questionnaires. For example, often investigators restrict the workplace exposure information collected to jobs, industries, and dates. From these limited data, they apply job exposure matrices to assign occupational exposure estimates. When applying the matrix, individuals holding a job are considered non-exposed if the exposure occurs only in a small proportion of workers in the job. This procedure will, however, inevitably result in classifying among the non-exposed individuals, a small proportion of workers who are, in reality, exposed. Similarly, individuals having jobs entailing a high probability of exposure will be considered exposed, even if they belong to the small proportion of non-exposed workers on this job. Detailed descriptions of tasks and work conditions of the jobs held by individual study subjects and evaluation of these data on the individual subject level are necessary for a better assessment. Thus, limited exposure data can contribute to misclassification, in that the available data (from which exposure is characterized) may not be representative of the individual's actual exposure level. This problem is more related to selection bias, is a general problem in epidemiology, and is not unique to exposure variables. The concept of bias is treated in chapter ▶Design and Planning of Epidemiological Studies of this handbook.

One can often recognize the circumstances in which differential misclassification may occur. Diseased subjects may have reflected more on their past exposures than the non-diseased (recall bias) or may take more care in providing correct responses. Differential bias will potentially occur when the exposure measurement instrument uses a human intermediary (e.g., the subject himself and/or an interviewer) aware of (or thinks he/she is aware of) the disease status. Thus, face-to-face interviews involve a substantial risk of producing interviewer effects. If a bias results from a different attitude of the interviewer toward the diseased compared with the non-diseased subjects, it is called interviewer bias. Self-administered questionnaires are generally believed to be less vulnerable to influences of response bias; however, the appearance of the questionnaire, the introductory letter, and the research group may all have an impact on response. The likelihood of bias from telephone interviews falls between these two data collection methods. Computer-assisted telephone interviewing has become the method of choice in many studies and often has a high response rate and few missing data (Nybo Andersen and Olsen 2002).

16.5.3 Quantification of Measurement Errors: Reproducibility Studies

Evaluation of the reproducibility of measurement instruments can be done by comparing the same instrument under the same conditions over time or by comparing various instruments under the same conditions at the same time. An example of the first type of study evaluated the reproducibility of a self-administered lifetime physical activity questionnaire (Chasan-Taber et al. 2002). Subjects reconstructed physical activity at four ages, starting at menarche, twice in the same mail questionnaire administered 1 year apart. All intraclass correlation coefficients used to measure reproducibility ranged from 0.78 to 0.87, with a value of 0.83 for total lifetime estimate of exposure.

The area of nutritional epidemiology (cf. chapter ▶Nutritional Epidemiology of this handbook) is one in which the design of proper questionnaire instruments has been extensively investigated. Subar et al. (2001) compared a new food frequency questionnaire and two widely used dietary questionnaires using telephone 24-h recalls. Despite substantial differences in the length and the design of the questionnaires, correlations obtained for dietary composition (i.e., total energy intake and 26 nutrients) were very similar. This comparison provides evidence that carefully designed self-administered food frequency questionnaires can provide reasonably reproducible measures of current nutrient intakes in epidemiological applications. There are still, however, questions about the validity of these instruments, and probably only the comparison of questionnaires with a truly uncorrelated error, such as a biochemical indicator of diet, will resolve these validity issues.

16.5.4 Quantification of Measurement Errors: Validation Studies

Ideally, an instrument should be evaluated by comparing it to a standard under the conditions the instrument is used. In evaluating the validity of any measurement instrument, the choice of the gold standard is a critical issue. Biochemical indicators of internal exposure provide an independent assessment for which measurement errors are not likely to be correlated with errors in air or water measurements or questionnaires. Biological measurements may represent historical exposures only if the chemical of interest has a sufficiently long biological half-life and may represent recent exposures only if the chemical has a relatively short half-life. In both cases, for the biological measure to be useful, the body burden cannot be affected by the disease or its treatment. In other situations, the biomarker may not measure the target agent of interest. Other challenges of biological monitoring can be found in Sect. 16.3.1 of this chapter. Biochemical indicators of dietary intake have a great appeal as the gold standard to assess the validity of dietary questionnaires (Willett 1990). There are limitations, however, in that the indicators may not reflect only dietary intake, and there are many dietary factors of interest for which there is no biomarker.

Practically, however, a gold standard often does not exist, especially when exposure has to be assessed retrospectively (e.g., historical tobacco consumption of individuals). For some exposures, a partial validation may be possible, by comparing questionnaire results to preexisting records. For example, reported jobs can be compared to employers' records, and smoking consumption can be compared to past medical records. Identification of gold standards that are alloyed and how to account for this error has been discussed (Wacholder et al. 1995). The validation of the instrument is also often measured by its ability to predict disease risk in prospective studies (Willett 1998). This approach is somewhat problematic, however, in that the epidemiological outcome is used to test the instrument. Nonetheless, a good instrument should produce better risk estimates than a poor one (Tielemans et al. 1998).

16.5.5 Methods for Correcting Measurement Errors

The effect of a systematic difference between the actual concentration and the concentration measured can be reduced or minimized simply by applying a correction factor reflecting the difference to the exposure estimate if the difference is known. Internal validation studies have been proposed to reduce the impact of measurement error. In one approach, exposure is measured, although imperfectly, from everyone in the study, and, simultaneously, a more accurate but more expensive measurement is collected on only a small subset of cases and controls selected randomly. Sophisticated statistical methods can then be applied in order to infer the corrected odds ratio from measurement error models fitted to the parallel exposure measurements from the validation sample (Stürmer et al. 2002). These so-called two-phase designs are among others investigated by Schill et al. (1993), Schill and Drescher (1997) and have been applied by Pohlabeln et al. (2002). This method, however, has not yet been routinely implemented, and further research is needed to establish the robustness of the procedures in realistic settings and to determine optimal designs for selecting a validation sample. As quoted by Chatterjee and Wacholder (2002) in a recent commentary, "The best way to reduce bias from measurement error is to improve tools for measuring exposures including biological markers, environmental samples and questionnaires."

A second approach that is gaining popularity is to conduct an uncertainty analysis (or sensitivity analysis; Rothman and Greenland (1998)). In this approach, investigators identify the uncertainty around a point estimate (e.g., two drinks of wine a day). For example, if the question "How many glasses of wine do you drink?" was asked and the responses were <1/day, 1–3/day, 4–5/day, and >5/day, the uncertainty ranges of these responses could be 0–0.9, 1–3, 4–5, and 6–10, respectively. Monte Carlo or other statistical simulations allow a better understanding of the uncertainty around the disease risk estimates.

16.6 Conclusions

The demand for accurate exposure assessment implies the need for development of validated and reliable tools in parallel with reduced costs and increased applicability in field studies. Sophisticated techniques are now available for direct measurement of chemicals in most mediums with excellent sensitivity and reproducibility. Similarly, questionnaires are being developed in various fields with considerable effort being put into their validation.

In some areas, such as occupational or environmental epidemiology, improvement is dependent upon additional knowledge on exposure determinants both at the personal and population levels and on objective comparisons of the quality of various available methods for exposure assessment (Liljelind et al. 2003). Quantitative estimates of exposure using statistical modeling are currently being developed, mainly for risk assessment purposes, but their applicability to epidemiological studies has not been fully explored.

To solve the problem of mixed exposures, the trend is toward building exposure indices summarizing several exposures according to biological hypotheses about their joint mechanisms of action. In the near future, new biotechnologies (e.g., genomics, proteomics) will contribute to the development of biomarkers of gene expression, intermediate between markers of exposure and markers of early effects that will summarize the joint action of mixed exposures at the molecular level (Henry et al. 2002; cf. chapter ▸Molecular Epidemiology of this handbook). The applicability of these techniques in epidemiological studies opens a whole new area of research.

References

Ahlborg UG, Becking GC, Birnbaum LS, Brouwer A, Derks HJGM, Feeley M, Golor G, Hanberg A, Larsen JC, Liem AKD, Safe SH, Schlatter C, Waern F, Younes M, Yrjänheikki E (1994) Toxic equivalency factors for dioxin-like PCBs. Chemosphere 28:1049–1067

Ainsworth BE, Richardson MT, Jacobs DR Jr, Leon AS, Sternfeld B (1999) Accuracy of recall of occupational physical activity by questionnaire. J Clin Epidemiol 52:219–227

Armstrong BK, White E, Saracci R (1994) Principles of exposure measurement in epidemiology. Monographs in epidemiology and biostatistics, vol 21. Oxford University Press, Oxford

Benke G, Sim M, Fritschi L, Alfred G, Forbes A, Kauppinen T (2001) Comparison of occupational exposure using three different methods: hygiene panel, job exposure matrix (JEM) and self reports. Appl Occup Environ Hyg 16:84–91

Blair A, Zahm SH (1993) Patterns of pesticide use among farmers: implications for epidemiologic research. Epidemiology 4:55–62

Bradburn NM, Rips LJ, Shevell SK (1987) Answering autobiographical questions: the impact of memory and inference on surveys. Science 236:157–161

Brunekreef B, Noy D, Clausing P (1987) Variability of exposure measurements in environmental epidemiology. Am J Epidemiol 125:892–898

Burstyn I, Kromhout H, Kauppinen T, Heikkila P, Boffetta P (2000) Statistical modelling of the determinants of historical exposure to bitumen and polycyclic aromatic hydrocarbons among paving workers. Ann Occup Hyg 44:54–56

Carpenter DO, Arcaro K, Spink DC (2002) Understanding the human health effects of chemical mixtures. Environ Health Perspect 110(suppl 1):25–42

Chasan-Taber L, Erickson JB, McBride JW, Nasca PC, Chasan-Taber S, Freedson PS (2002) Reproducibility of a self-administered lifetime physical activity questionnaire among female college alumnae. Am J Epidemiol 155:282–289

Chatterjee N, Wacholder S (2002) Validation studies: bias, efficiency and exposure assessment. Epidemiology 13:503–506

Cherrie JW, Schneider T, Spankie S, Quinn M (1996) A new method for structured, subjective assessments of past concentrations. Occup Hyg 3:75–83

Gladen BC, Doucet J, Hansen LG (2003) Assessing human polychlorinated biphenyl contamination for epidemiologic studies: lessons from patterns of congener concentrations in Canadians in 1992. Environ Health Perspect 111:437–443

Grandjean P, Bjerve KS, Weihe P, Steuerwald U (2001) Birthweight in a fishing community: significance of essential fatty acids and marine food contaminants. Int J Epidemiol 30:1272–1278

Heederik D, Attfield M (2000) Characterization of dust exposure for the study of chronic occupational lung disease: a comparison of different exposure assessment strategies. Am J Epidemiol 151:982–990

Henry CJ, Phillips R, Carpanini F, Corton JC, Craig K, Igarashi K, Leboeuf R, Marchant G, Osborn K, Pennie WD, Smith LL, Teta MJ, Vu V (2002) Use of genomics in toxicology and epidemiology: findings and recommendations of a workshop. Environ Health Perspect 110:1047–1050

Hines CJ, Selvin S, Samuels SJ, Hammond SK, Woskie SR, Hallock MF, Schenker MB (1995) Hierarchical cluster analysis for exposure assessment of workers in the semiconductor health study. Am J Ind Med 28:713–722

Hsing AW (2001) Hormones and prostate cancer: what's next. Epidemiol Rev 23:42–58

Kriebel D, Checkoway H, Pearce N (2007) Exposure and dose modelling in occupational epidemiology. Occup Environ Med 64:492–498

Liljelind I, Rappaport S, Eriksson K, Andersson J, Bergdahl IA, Sunesson AL, Jarvholm B (2003) Exposure assessment of monoterpenes and styrene: a comparison of air sampling and biomonitoring. Occup Environ Med 60:599–603

Lindzey G, Aronson E (1985) Handbook of social psychology, vol I, 3rd edn. Random House, New York

Longnecker MP, Rogan WJ, Lucier G (1997) The human health effects of DDT (dichlorodiphenyltrichloroethane) and PCBs (polychlorinated biphenyls) and an overview of organochlorines in public health. Annu Rev Public Health 18:211–244

McGrail MP, Stewart W, Schwartz BS (1995) Predictors of blood lead levels in organolead manufacturing workers. J Occup Environ Med 37:1224–1229

Nybo Andersen A-M, Olsen J (2002) Do interviewer's health beliefs and habits modify responses to sensitive questions? A study using data collected from pregnant women by means of computer-assisted telephone interviews. Am J Epidemiol 155:95–100

Pohlabeln H, Wild P, Schill W, Ahrens W, Jahn I, Bolm-Audorff U, Jöckel K-H (2002) Asbestos fibreyears and lung cancer: a two-phase case-control study with expert exposure assessment. Occup Environ Med 59:410–414

Reif JS, Burch JB, Nuckols JR, Metzger L, Ellington D, Anger WK (2003) Neurobehavioral effects of exposure to trichloroethylene through a municipal water supply. Environ Res 93:248–258

Rothman KJ, Greenland S (eds) (1998) Modern epidemiology, 2nd edn. Lippincott-Raven, Philadelphia, pp 115–134

Samet JM (1995) What can we expect from epidemiologic studies of chemical mixtures? Toxicology 105:307–314

Schantz SL, Widholm JJ, Rice DB (2003) Effects of PCB exposure on neuropsychological function in children. Environ Health Perspect 111:357–376

Schill W, Drescher K (1997) The analysis of case-control studies under validation subsampling: a comparison of four approaches. Stat Med 16:117–132

Schill W, Jöckel K-H, Drescher K, Timm J (1993) Logistic analysis in case-control studies under validation sampling. Biometrika 80:339–352

Soto AM, Fernandez MF, Luizzi MF, Oles-Karasko AS, Sonnenschein C (1997) Developing a marker of exposure to xenoestrogen mixtures in human serum. Environ Health Perspect 105(suppl 3):647–654

Stürmer T, Thürigen D, Spiegelman D, Blettner M, Brenner H (2002) The performance of methods for correcting measurement error in case-control studies. Epidemiology 13:507–516

Subar AF, Thompson FE, Smith AF, Jobe JB, Ziegler RG, Potischman N, Schatzkin A, Hartman A, Swanson C, Kruse L, Hayes RB, Lewis DR, Harlan LC (1995) Improving food frequency questionnaires: A qualitative approach using cognitive interviewing. J Am Diet Assoc 95: 781–788

Subar AF, Thompson FE, Kipnis V, Midthune D, Hurwitz P, McNutt S, McIntosh A, Rosenfeld S (2001) Comparative validation of the Block, Willett, and National Cancer Institute Food Frequency Questionnaires. The Eating at America's Table Study. Am J Epidemiol 154:1089–1099

Sudman S, Bradburn NM (1982) Asking questions. Jossey-Bass Publishers, San Francisco

Tielemans E, Heederik D, Burdorf A, Vermeulen R, Veulemans H, Komhout H, Hartog K (1998) Assessment of occupational exposures in a general population: comparison of different methods. Occup Environ Med 56:145–151

van der Woord MP, Kromhout H, Barregard L, Jonsson P (1999) Within-day variability of magnetic fields among electric utility workers: consequences for measurement strategies. Am Ind Hyg Assoc J 60:713–719

Vermeulen R, Stewart P, Kromhout H (2002) Dermal exposure assessment in occupational epidemiologic research. Scand J Work Environ Health 28:371–385

Wacholder S, Hartge P, Dosemeci M, Armstrong B (1995) Validation studies using an alloyed gold standard (letter). Am J Epidemiol 141:277

Willett WC (1998) Invited commentary: comparison of food frequency questionnaires. Am J Epidemiol 148:1157–1159

Willett WC (1990) Reproducibility and validity of food-frequency questionnaires. In: Nutritional epidemiology. Oxford University Press, New York

Wu J, Wilhelm M, Chung J, Ritz B (2011) Comparing exposure assessment methods for traffic-related air pollution in an adverse pregnancy outcome study. Environ Res 111:685–692

Misclassification 17

Paul Gustafson and Sander Greenland

Contents

P. Gustafson (✉)
Department of Statistics, University of British Columbia, Vancouver, BC, Canada

S. Greenland
Department of Epidemiology, School of Public Health, University of California, Los Angeles, CA, USA

Department of Statistics, UCLA College of Letters and Science, Los Angeles, CA, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_58,

17.1 Introduction

The convention in epidemiology and biostatistics is to divide the study of mismeasured variables into the areas of measurement error for continuous variables and misclassification for categorical variables. Although the topics overlap considerably, chapter ▸Measurement Error of this handbook focuses on measurement error, whereas the present chapter is devoted to misclassification. As a motivating example of a misclassified variable in an epidemiological study, say that a binary exposure is ascertained via subject self-report on a questionnaire. Given human memory limitations, we would usually expect a portion of responses to be erroneous. For instance, in the study of Kraus et al. (1989) on possible association between maternal antibiotic use during pregnancy and sudden infant death syndrome (SIDS), antibiotic use is self-reported by subjects via questionnaire. Examination of medical records of some subjects, however, indicates that the questionnaire responses are erroneous for some subjects. Thus, antibiotic use as determined via questionnaire is subject to misclassification. Moreover, this misclassification has implications when the association between antibiotic use and SIDS is inferred.

As another example, Natarajan (2009) considers whether carotenoid intake beyond a certain threshold correlates with blood pressure. Carotenoid intake is measured indirectly from dietary recall interviews conducted by nutritionists, and hence is subject to error both in recall and in the conversion from diet to nutrient data. Whereas carotenoid intake itself is a continuous variable subject to measurement error, the problem becomes one of misclassification when the focus shifts to the binary indicator of whether intake exceeds the threshold.

This chapter covers some of the main ideas and issues in dealing with misclassified variables in epidemiological contexts. Section 17.2 focuses on the impact of not dealing with misclassification, i.e., what happens if we proceed to analyze the data as if, contrary to fact, the exposure is measured correctly on all subjects. This starts with a basic case in Sect. 17.2.1, involving only a binary exposure and outcome along with a simple form of misclassification. More complex situations are then considered in Sect. 17.2.2. Section 17.3 turns to analyzing data while acknowledging the misclassification at play. After describing a simple situation in Sect. 17.3.1, the issue of where information about misclassification comes from is dealt with in Sect. 17.3.2. Then, different inferential strategies are contrasted in Sect. 17.3.3. Section 17.4 deals briefly with outcome and confounding variables subject to misclassification. Some summary remarks are given in Sect. 17.5.

17.2 The Impact of Uncontrolled Misclassification

17.2.1 The Basic Case

Consider a binary exposure X, with interest in how X is associated with a disease indicator Y that is correctly ascertained for all study subjects. *Exposure* *misclassification* results if the ascertained exposure W differs from the true exposure X for some subjects. The accuracy of the classification procedure is usually characterized by the sensitivities $Sn_{\mathrm{y}} = \Pr(W = 1|X = 1, Y = y)$ and specificities $Sp_{\mathrm{y}} = \Pr(W = 0|X = 0, Y = y)$; these are the probabilities that the ascertainment is correct for a given subject. These W probabilities should be contrasted with the X probabilities showing the correctness of the classification method when applied to a given population, the positive predictive values $\Pr(X = 1|W = 1, Y = y)$, and the negative predictive values $\Pr(X = 0|W = 0, Y = y)$. The predictive values can be derived from Sn_{y}, Sp_{y}, and the true prevalence $p = \Pr(X = 1)$ via Bayes theorem,

$$\Pr(X = 1|W = 1, Y = y) = Sn_{\mathrm{y}} \cdot p/[Sn_{\mathrm{y}} \cdot p + (1 - Sp_{\mathrm{y}}) \cdot (1 - p)] \quad \text{and}$$
$$\Pr(X = 0|W = 0, Y = y) = Sp_{\mathrm{y}} \cdot (1 - p)/[(1 - Sn_{\mathrm{y}}) \cdot p + Sp_{\mathrm{y}} \cdot (1 - p)].$$

It is important to note that our intuition tends to be poor in translating between classification probabilities and predictive values. As an example, say $Sn_0 = 0.95$, $Sn_1 = 0.90$, $Sp_0 = 0.88$, and $Sp_1 = 0.92$. These classification probabilities seem generally high, and we might characterize the misclassification as being somewhat modest. If, however, exposure is rare, the positive predictive values can be quite low. For instance, if $p = 0.1$, then Bayes theorem yields the positive predictive values as 0.47 (for $Y = 0$) and 0.55 (for $Y = 1$), whereas the negative predictive values are very high: 0.994 (for $Y = 0$) and 0.988 (for $Y = 1$).

Knowing the extent of misclassification alone does not suffice to characterize its impact on statistical inference since various constraints may apply. If only exposure is misclassified, the simplest constraint is that of *non-differential* misclassification, whereby the ascertainment of exposure status is completely independent of other variables. In the present setting, this is expressed as conditional independence between W and Y given X, i.e., $\Pr(W = w|X = x, Y = y) = \Pr(W = w|X = x)$, so that we may drop the "y" subscript from Sn_{y} and Sp_{y}. More involved settings would replace Y with "Y and other variables" in this conditional-independence statement.

On first glance, non-differential misclassification appears to be realistic whenever the ascertainment of W is carried out without knowledge of Y. Unfortunately, the situation is often not so simple, as discussed below. Nonetheless, the non-differential case is the typical starting point for analysis of misclassification effects since even with the constraint, we obtain an association between observables W and Y which differs from the association of real interest, between X and Y. Specific equations for describing and accounting for this phenomenon are given, for instance, in Sect. 3.3 of Gustafson (2003) and p. 356 of Greenland and Lash (2008). For binary variables with only exposure misclassified, the association between W and Y is always weaker than that between X and Y, resulting in the often-quoted maxim that exposure measurement error or misclassification biases results toward the null (i.e., pushes results toward no association). This maxim is often quoted outside of its rather limited domain of applicability, as will be discussed later on.

Table 17.1 Population odds ratio for misclassified exposure and disease, for various values of the odds ratio for actual exposure and disease, the exposure prevalence among non-diseased, and the sensitivity and specificity of the non-differential misclassification

		$OR = 1.25$		$OR = 2.0$	
Prevalence	Sp	$Sn = 0.95$	$Sn = 0.80$	$Sn = 0.95$	$Sn = 0.80$
0.01	0.95	1.04	1.03	1.16	1.14
	0.80	1.01	1.01	1.05	1.04
0.10	0.95	1.17	1.15	1.67	1.60
	0.80	1.08	1.07	1.34	1.27
0.20	0.95	1.20	1.19	1.80	1.71
	0.80	1.13	1.11	1.52	1.42

To provide some examples, Table 17.1 gives the population odds ratio between W and Y resulting from particular values for (i) the population prevalence of X among the non-diseased, (ii) the population odds ratio between X and Y, and (iii) the sensitivity and specificity of the non-differential misclassification. The attenuation of the odds ratio toward the null is seen to be more dramatic when either exposure ($X = 1$) or non-exposure ($X = 0$) is uncommon. For exposure prevalence below 50%, attenuation is seen to be driven more by imperfect specificity than by imperfect sensitivity; the reverse is true if exposure prevalence is over 50%. In practical terms, the discrepancy between the (W, Y) and (X, Y) associations can be very substantial under realistic values for exposure prevalence and misclassification magnitude. Thus, simply treating W as "the exposure" can be highly misleading. Valid estimation of the (W, Y) association will typically be highly invalid for the (X, Y) association.

17.2.2 More Complex Situations

As alluded to above, the "bias toward the null" maxim is often violated. It can be derived under special conditions such as non-differential misclassification of a binary exposure alone (Newell 1962) and generalizes to quantitative exposures under certain regression constraints (Weinberg et al. 1994). But in situations where exposure is not truly binary, errors depend on other variables, or data are grouped (ecologic), bias toward the null is no longer guaranteed even if non-differentiality holds (Walker and Blettner 1985; Dosemeci et al. 1990; Birkett 1992; Chavance et al. 1992; Kristensen 1992; Brenner et al. 1992; Weinberg et al. 1994).

To illustrate that non-differential misclassification does not always induce bias toward the null, consider a trinary exposure X with levels labeled "none," "low," and "high." Under non-differential misclassification, this gives rise to apparent exposure W, where this misclassification is described by a 3 by 3 matrix of probabilities of W given X. For a given misclassification matrix and a given joint distribution of actual exposure X and disease Y, it is a straightforward probability calculation to determine the joint distribution of apparent exposure W and disease Y.

Table 17.2 The effect of non-differential misclassification in a trinary exposure example. The left side of the table gives a joint distribution of actual exposure and disease, (X, Y). For the misclassification probabilities given in the text, the right side of the table gives the resulting joint distribution of apparent exposure and *disease* (W, Y)

	X			W		
	No	Low	High	No	Low	High
$Y = 1$	0.1	0.1	0.1	0.095	0.110	0.095
$Y = 0$	0.3	0.3	0.1	0.265	0.290	0.145

As a simple example, say that for each level x of X the probability of correct classification is $\Pr(W = x|X = x) = 0.7$. For the low level of X, suppose the probabilities of the two possible misclassifications (to none or to high) are 0.15 each. For both the none and high levels of X, suppose the probability of classification into the adjacent (low) category is 0.2, while the probability of classification into the distant category is 0.1. Thus, the probability of correct classification exceeds the probability being off by one level, which in turn exceeds the probability of being off by two levels. When applied to the joint distribution of (X, Y) given in the left side of Table 17.2, this non-differential misclassification yields the joint distribution of (W, Y) in the right side of this table. Note that the (X, Y) association in this example is described by a disease odds ratio of 1, contrasting the low level of X to none, and a disease odds ratio of 3 contrasting the high level of X to the low level. However, the (W, Y) association has odds ratios of 1.06 contrasting low to none and 1.73 contrasting high to low levels. Thus, we see slight bias *away* from the null in the first instance and (severe) bias toward the null in the second instance. This example underscores the lack of a simple general rule concerning the impact of non-differential misclassification when there are more than two levels of the exposure.

Another setting where often-seen rules do not apply is that of differential misclassification. In fact, for any given distribution of binary exposure X and binary outcome Y, one can find Y-dependent misclassification probabilities (sensitivities and specificities) for the misclassification from X to W which give any answer desired for the (W, Y) association. This in itself is not surprising. What is perhaps more surprising, however, is the frail nature of the non-differential assumption. Intuitively, non-differential would seem to say that the process by which X leads to W is blind to Y. However, many situations do not lend themselves to such blinding; moreover, some situations which appear on first glance to involve blind classification are seen upon closer inspection to be differential (Flegal et al. 1991; Wacholder et al. 1991).

Studies with retrospective ascertainment of exposures (e.g., questions based on recall of past events) are a prime example of settings where non-differentiality may be most dubious. This is particularly true when W arises as a response on a self-report questionnaire administered after diagnosis since subjects are necessarily not blinded to their own disease status. One can envision situations where a psychological tendency to blame a putative exposure for illness might induce non-exposed cases to declare themselves exposed more easily than non-exposed non-cases (so case specificity is lower than non-case specificity) or in which the awareness of the illness sharpens recall efforts and induces exposed cases to remember actual exposures at a higher rate than exposed non-cases (so case sensitivity is higher than non-case sensitivity).

On the other hand, in "denial" settings where exposure involves a behavioral choice that might be stigmatized, exposed cases might be particularly hesitant to admit that a behavior could possibly have contributed to their disease (so case sensitivity is lower than non-case sensitivity), or exposed cases might altruistically or fatalistically be less inhibited in admitting to these exposures than non-cases (so case sensitivity is higher than non-case sensitivity). Thus, retrospective ascertainment may be particularly fertile grounds for engendering differential misclassification, although psychological speculations may provide little guidance as to the direction or net impact of such differentiality (Drews and Greenland 1990).

Even when the misclassification process is physically independent of disease status, there may be mechanisms by which the non-differential assumption fails to hold.

As an example, suppose that the binary exposure X arises from thresholding a continuous exposure variable X_c. Similarly, the apparent exposure W might arise from thresholding W_c, which is a noisy measurement of X_c. A first intuition might be that if the process by which X_c gives rise to W_c is blind to Y, then the thresholded quantities would behave similarly. That is, non-differential measurement error of X_c might be thought to induce non-differential misclassification of X. Unfortunately, however, the situation is more complicated. If Y depends on X_c not just through X, then generally W and Y will not be conditionally independent given X. That is, if there is still a risk gradient with X even when conditioning on X being on a particular side of the threshold, then non-differential error in W_c as a measure of X_c gives rise to differential misclassification in W as a measure of X. This is then an example of an outward appearance of non-differential misclassification obscuring truth to the contrary.

For further discussion of misclassification arising from thresholding a poorly measured continuous variable, see Flegal et al. (1991), Gustafson and Le (2002) and Natarajan (2009). The latter article gives an example where the binary predictor X is an indicator of the log plasma carotenoid level exceeding a threshold (chosen to be one population standard deviation above the population mean level). Misclassification then arises because the measured carotenoid concentration W_c involves error relative to the actual level X_c.

17.3 Accounting for Misclassification

17.3.1 A Basic Case

Having discussed the manner in which the (X, Y) association for the actual exposure and disease can differ from the (W, Y) association for the assessed exposure and disease, we now turn to the issue of inferring the (X, Y) relationship from (W, Y) data. The feasibility of doing this is strongly governed by the form and extent of information about the misclassification. Simply put, moving from X to W corresponds to a distortion of the data, and we cannot expect to undo the impact of this distortion without some information on how badly W represents X.

As a simple but admittedly unrealistic starting point, if we have misclassification of a binary X into a binary W with known sensitivities and specificities, then inferential procedures are straightforward. In this case, the prevalences of apparent exposure, $\Pr(W = 1|Y = y)$ for $Y = 0, 1$, and the prevalences of true exposure, $\Pr(X = 1|Y = y)$, are completely determined by the system of equations

$$\Pr(W = 1|Y = y) = \Pr(X = 1|Y = y)Sn_{\mathrm{y}} + \Pr(X = 0|Y = y)(1 - Sp_{\mathrm{y}}). \quad (17.1)$$

Given Sn_{y}, Sp_{y}, and data on $(W|Y)$, (17.1) represents two equations (one for each Y value) in two unknowns $\Pr(X = 1|Y = 1)$ and $\Pr(X = 1|Y = 0)$. These equations can be solved explicitly, leading to odds ratio estimates relating X and Y. Considerable caution is needed, however, when applying this system with constraints, e.g., assuming non-differentiality (i.e., using the same values for Sn_{y} and Sp_{y} for cases and non-cases), as this can lead to estimates more biased than those from the uncorrected data (Brenner 1996). Fortunately, system (17.1) allows differential misclassification and generalizes straightforwardly to polytomous exposures and outcomes as

$$\Pr(W = w|Y = y) = \sum_{\mathrm{x}} \Pr(X = x|Y = y)\Pr(W = w|X = x, Y = y),$$

which can be solved for $\Pr(X = x|Y = y)$ via ordinary matrix inversion (Greenland and Kleinbaum 1983; Greenland and Lash 2008).

More generally, but with the same caveats, weighted least squares, maximum likelihood and Bayesian methods are readily applied to convert inferences about the (W, Y) relationship to inferences about the (X, Y) relationship. A technical difficulty for frequentist methods based on Eq. 17.1 however is that sampling variation can result in a sample prevalence of apparent exposure $(W = 1)$ which maps to a prevalence of actual exposure $(X = 1)$ that is less than zero or greater than one (Tu et al. 1994, 1995; Greenland and Lash 2008). As an example, suppose that classification probabilities $Sn_0 = Sn_1 = 0.8$ and $Sp_0 = Sp_1 = 0.85$ are known to the investigator. However, the investigator is unaware that $\Pr(X = 1|Y = 0) = 0.05$ and $\Pr(X = 1|Y = 1) = 0.10$, and consequently $\Pr(W = 1|Y = 0) = 0.183$ and $\Pr(W = 1|Y = 1) = 0.215$. Now say data in the form of W measurements on 200 controls and 200 cases are available. At these sample sizes, sample prevalences for W are likely within ±0.05 of the population prevalences. For instance, say the sample estimates for $\Pr(W = 1|Y = y)$ are 0.14 (for $y = 0$) and 0.22 (for $y = 1$). Now applying Eq. 17.1 in the other direction, our corresponding estimates for $\Pr(X = 1|Y = y)$ would be -0.016 and 0.108, for $Y = 0$ and $Y = 1$, respectively. The former (negative) estimate is clearly non-sensical and does not lead to a sensible estimate for the (X, Y) association. Gustafson et al. (2001) demonstrate that formal Bayesian analyses obviate this concern.

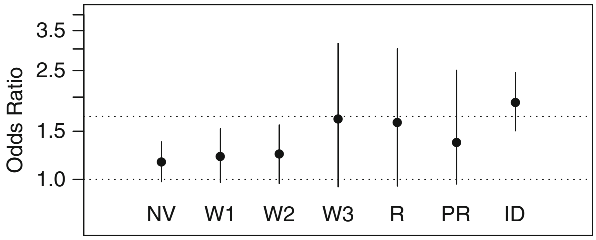


Fig. 17.1 Odds ratio inferences from a simulated dataset with 111 of 1,250 controls and 197 of 1,250 cases actually exposed, but 390 of the controls and 430 of the cases apparently exposed. Point and 95% interval estimates are displayed on a logarithmic scale. The ideal (ID) analysis uses the actual exposure data (X), while the other analyses use the apparent exposure data (W). The naive (NV) analysis simply treats W as if it were X. The "right" (R)-adjusted analysis correctly takes $Sn = 0.9$ and $Sp = 0.75$ as known. The three "wrong"-adjusted analyses ($W1$ through $W3$) incorrectly take (Sn, Sp) as known, with values (0.9, 0.9), (0.75, 0.9), and (0.75, 0.75), respectively. The adjustment based on a prior distribution (PR) for (Sn, Sp) uses the prior distribution described in the text. The *dotted lines* indicate the null and true values

Perfect knowledge of the sensitivity and specificity of exposure misclassification is unavailable in real problems. More realistically, one might hope to have some *validation data* available from which the relation between W and X can be estimated. This situation is discussed below. Absent such data though, one might have to rely upon more amorphous information, such as investigator-provided ranges of plausible values for the sensitivity and specificity of exposure classification. This puts the problem into the realm of sensitivity analysis, examining a range of answers corresponding to a range of assumptions about the severity of misclassification. Rather than provide hard constraints in the form of intervals for Sn_{y} and Sp_{y}, one might advocate softer, probabilistic constraints in the form of prior distributions for these quantities. That is, inference is still built upon (17.1), but with the acknowledgement that Sn_{y} and Sp_{y} are only known with some degree of uncertainty or vagueness, rather than exactly.

As an example, consider a population with an odds ratio between X and Y of 1.7, with 10% exposure prevalence among the non-diseased. A computer-simulated sample of 1,250 controls and 1,250 cases from the population produced 111 truly exposed controls and 197 truly exposed cases. However (again via computer simulation), non-differential misclassification with sensitivity $Sn = 90\%$ and specificity $Sp = 75\%$ resulted in the investigator observing 390 apparently exposed controls and 430 apparently exposed cases.

Point and interval estimates for the (X, Y) odds ratio arising from various Bayesian analyses of these simulated data are depicted in Fig. 17.1. These include the ideal (but unattainable in real settings) analysis arising from the (X, Y) data, and the naïve analysis arising from the (W, Y) data without any misclassification adjustment. Five adjusted analyses based on the (W, Y) data are also presented. Four of these are based on taking the (Sn, Sp) values as known (correctly in the first instance and incorrectly in the other three). A fifth adjusted analysis is based on a prior distribution for (*Sn, Sp*) which subsumes the different values specified. Particularly, this prior distribution has 5th and 95th percentiles for both *Sn* and *Sp* as 0.75 and 0.90. More technically, *Sn* and *Sp* are assumed independent and identically distributed *a priori*, each with a Beta (50.4, 10.2) distribution. This specification is somewhat unrealistic insofar as various forces may induce dependencies between *Sn* and *Sp*, but since these forces may be in either direction, independence is a simple starting point. Similarly, contextually realistic priors for the underlying (X, Y) distribution would be based on independent distributions for the odds ratio and the control exposure $(X = 1)$ prevalence; see Greenland (2001) and Greenland and Lash (2008, p. 369) for discussion. But, for the sake of simplicity, our analyses will use independent uniform distributions as priors for the case and control exposure prevalences.

In this example, we see that the inability to measure X precisely comes with a steep cost. The naïve analysis is biased, with a 95% interval estimate missing the true value. The adjusted analyses which use good information about misclassification (either correct knowledge of the values of *Sn* and *Sp*, or prior distributions that concentrate on sufficiently narrow ranges of values that include the correct values) are helpful in appropriately moving the naïve point estimate away from the null, and yielding interval estimates compatible with the true value. However, the widths of these intervals, roughly double that of the ideal analysis, show the severe loss of information entailed by the misclassification. Worse, in this particular case, the intervals cover the null, illustrating how misclassification induces a loss of power for detecting associations.

To emphasize that the behavior seen in Fig. 17.1 is typical in the 2 by 2 case, results for a further nine datasets, simulated under the same conditions described above, appear in Fig. 17.2. (To conserve space, only one of the three analyses with wrong values is presented.) This gives a slightly more nuanced view of how much power and precision is lost upon replacing X with W, even if the extent of misclassification is correctly known. Interestingly, in aggregate the interval estimates arising from the correctly known values of (*Sn, Sp*) appear to be slightly longer than those arising from the prior distribution applied to (*Sn, Sp*), i.e., less uncertainty *a priori* transmutes to relatively more uncertainty *a posteriori*. Gustafson and Greenland (2006) take up this curious issue in some depth.

Another related point is that the naïve analysis and the analysis using the correct values coincide in terms of which datasets yield interval estimates excluding the null. Thus, while adjustment for non-differential misclassification can be helpful for estimation purposes, it seems to have no impact for testing the null hypothesis other than loss of power from using the uncorrected (W, Y) data (Bross 1954). Greenland and Gustafson (2006) stress that this is a general finding for likelihood-ratio hypothesis testing under non-differential exposure misclassification, although Gustafson and Greenland (2006) show that the situation can be more complicated for Bayesian analysis.

It is worth noting that the severe bias from misclassification seen in the 2×2 case can be considerably worsened in matched analyses (Greenland 1982) and can lead to spurious appearance or masking of heterogeneity (effect-measure modification)

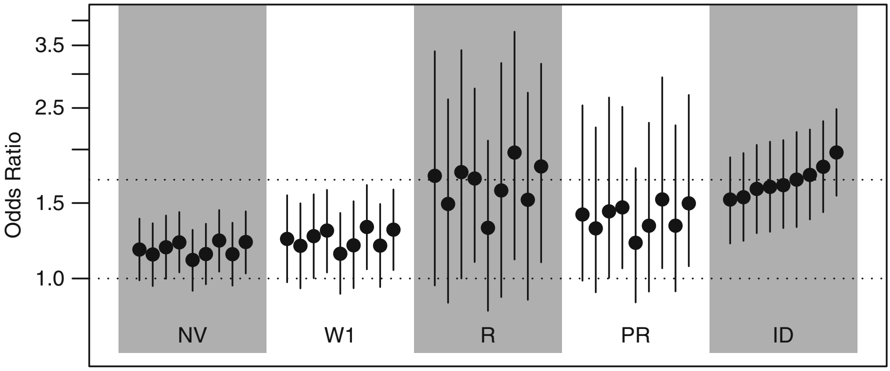


Fig. 17.2 Odds ratio inference for nine further synthetic datasets generated under the same conditions as in Fig. 17.1. These datasets are deliberately ordered to have ascending sample odds ratios for (X, Y)

in stratified analyses (Greenland 1980). Finally, as alluded to earlier, the non-differentiality assumption is more easily violated in practice than often realized (Wacholder et al. 1991), and its violation can play havoc with methods that assume it holds (Brenner 1996).

17.3.2 Sources of Information About Misclassification

A key concern in misclassification problems is how information about the extent of misclassification is derived. The previous subsection dealt with the somewhat amorphous case where the investigator has rough speculations about the extent of misclassification, which can be cast in the form of a prior distribution. In other cases, however, the data at hand may speak directly to the misclassification magnitude.

17.3.2.1 Validation Data

In some settings, it may be possible to measure the true exposure X, but these measurements are too expensive to make for all study subjects. This leads to the notion of *validation data*. Internal validation arises if the (W, Y) measurements for all study subjects are augmented with X measurements, ideally for a randomly selected subset of the subjects. The random selection could be undertaken as a random sample of all subjects, or it would also be permissible and could be more efficient to use stratified sampling with strata defined by the levels of W and Y. As an example of study data with a validation component, Kraus et al. (1989) consider antibiotic prescription during pregnancy (X) and sudden infant death syndrome (Y), where the noisy assessment of exposure (W) is self-reported antibiotic use on a questionnaire, recorded for all 797 controls and 775 cases.

Medical records were examined for 217 controls and 211 cases. Presuming these records to be accurate and complete, X itself is thus observed for these subjects.

Randomly sampled internal validation data lead to identified statistical inference – as the overall study sample size grows one learns the relationship between W and Y more precisely, as the size of the validation subset grows one learns the relationship between W and X given Y more precisely, and (if as here sampling is random within Y levels) one also learns directly about the odds ratio relating X to Y.

Conceptually, we can extract all the information from this form of data by using sample proportions based on *all* subjects with $Y = y$ to estimate $\Pr(W = w|Y = y)$ and using sample proportions based on *validated* subjects with $W = w, Y = y$ to estimate $\Pr(X = x|W = w, Y = y)$, i.e., to estimate the predictive values. Since

$$\Pr(X = 1|Y = y) = \sum_{\mathrm{w}} \Pr(X = 1|W = w, Y = y)\Pr(W = w|Y = y), \quad (17.2)$$

our estimates of the quantities on the right-hand side of the equation yield estimates of $\Pr(X = 1|Y = y)$ for y = 0, 1, and hence we can infer the (X, Y) association (Marshall 1990; Brenner and Gefeller 1993; Lyles 2002).

To illustrate with the data from Kraus et al. (1989), we estimate $\Pr(W=1|Y=y)$ as 134/797 for $y = 0$ and 173/775 for $y = 1$, which, incidentally, gives the sample (W, Y) odds ratio as 1.42. Then, we estimate $\Pr(X = 1|W = w, Y = y)$ as 16/184 for $w = 0, y = 0$; 21/33 for $w = 1, y = 0$; 17/160 for $w = 0, y = 1$; and 29/51 for $w = 1, y = 1$. Plugging these into (17.2), we estimate $\Pr(X = 1|Y = 0)$ as 0.179, and $\Pr(X = 1|Y = 1)$ as 0.209, yielding an estimated (X, Y) odds ratio of 1.21. Note that this is an instance where adjustment for misclassification pushes the point estimate toward the null, as can indeed happen under differential misclassification.

For planning a study, however, if the accuracy of W as a measure of X is poor, the statistical contribution of the non-validated data may be low; thus the cost-effectiveness of using W at all instead of obtaining X on everyone may be called into question if the cost of obtaining X is not considerably more than the cost of obtaining W (Greenland 1988). Furthermore, the naïve use of adjustment methods that assume no error in X can leave considerable bias, perhaps even more than there was before adjustment (Wacholder et al. 1993), and that bias is worsened by incorrectly assuming non-differentiality (Brenner 1996).

At the other extreme, it may be impossible to obtain X on the study subjects, leaving external validation data as the only alternative to speculation. It may be unreasonable to expect the predictive values in Eq. 17.2 to transfer across populations. If however we have (X, W, Y) values for subjects outside the main study, we might reasonably hope that the $(W|X, Y)$ distribution is similar in both populations, in which case Eq. 17.1 or its generalizations can be applied. In situations where the non-differential misclassification assumption can be invoked, it would suffice to have measurements of (X, W) values for subjects from the external population, which might be limited to only cases or only non-cases.

17.3.2.2 Multiple Surrogates

In lieu of validation measurements, another potential source of information about misclassification arises from multiple surrogates. For instance, say that data on

(W_1, W_2, Y) are available, where W_1 and W_2 are two different surrogates for X. For instance, in the example mentioned above, the review of medical records is in fact imperfect in assessing maternal *use* of antibiotics (as opposed to mere prescription, i.e., intent-to-treat). Thus, if X is now defined as antibiotic use rather than prescription, both the questionnaire and the examination of medical records only yield surrogates for the actual X. Under some assumptions, data on multiple surrogates plus disease outcome can be informative for the relationship between X and Y, even though X is never observed. Arguably, however, the assumptions are strong. In addition to the assumption that each surrogate arises via non-differential misclassification of X, an assumption that the surrogates are conditionally independent of one another given X is required, or at least that the dependency structure is precisely known. This is typically hard to justify. In many settings, one can imagine that subjects with a higher risk of W_1 being in error could also be at particular risk of having W_2 be in error.

Despite this caveat, there is some intuition as to why two conditionally independent surrogates suffice. The basic setup requires the estimation of six quantities: exposure prevalences among controls and cases, plus sensitivity and specificity for each surrogate. Without the conditional-independence assumption, there would be several further parameters at play. However, the data-structure identifies the conditional distribution of $(W_1, W_2|Y)$, which indeed permits the estimation of (at most) six quantities (Hui and Walter 1980).

One can relax the assumption that the two surrogates are conditionally independent (see, for instance, Dendukuri and Joseph 2001; Hanson et al. 2003). These approaches necessarily involve a non-identified model and typically a Bayesian analysis with a suitable prior on the extent of departure from conditional independence. More than two surrogates can also be used (e.g., Chu et al. 2009). Some technical investigation of the identification issues for multiple surrogate models is given by Jones et al. (2010) and Gustafson (2009).

Much of the work on multiple surrogates arises in the context of diagnostic testing or screening, where it is not exposure per se, but rather disease status, that is subject to misclassification. Moreover, the focus may not be on relating disease status to other variables, but may rather lie in disease prevalence or the classification probabilities for disease-status surrogates. Overviews of the statistical aspects of diagnostic testing are given by Pepe (2003) and Zhou et al. (2002); see Broemeling (2007) for a more Bayesian discussion. While diagnostic testing is outside the purview of the present chapter, many of the ideas from this literature carry relevance for the situation of misclassified exposure.

17.3.3 Different Inferential Approaches

While the example of Sect. 17.3.1 uses Bayesian methods to draw inferences about the relationship between actual exposure and disease, the literature describes a wide array of methods applied to misclassification problems. One surprising feature of the literature is that many methodological suggestions fall outside the realm of likelihood-based methods and emphasize intuitive appeal. For instance, Küchenhoff et al. (2006) extend the *simulation-extrapolation* (SIMEX) procedure originally devised for continuous X (Cook and Stefanski 1995) to the discrete X setting. The appealing intuition here is that one can add varying degrees of extra misclassification to the already misclassified data, to establish the relationship between the (W, Y) association and the extent of misclassification. This relationship can then be extrapolated back to the no misclassification case, though in practice there can be uncertainty about the appropriate functional form to be used for this extrapolation.

In continuous variable settings, likelihood-based procedures (i.e., Bayesian and maximum likelihood procedures) are sometimes argued against on the grounds that the required model specification is too arduous and prone to misspecification. In discrete-data settings, however, such arguments are typically obviated because discrete variables do not entail extra assumptions about distributional forms. Thus, problems of misclassification for discrete variables seem prime candidates for likelihood-based methods, given the general large-sample efficiency that accompanies such methods. Greenland (2008) finds that traditional external validation estimators based on Eq. 17.1 are indeed maximum likelihood estimators, although they were not originally derived as such, and that some proposed estimators which are not likelihood-based can be quite inefficient relative to those which are.

Within the realm of likelihood-based procedures, there are several general issues. First, the difficulty of implementation varies considerably with problem. A primary example of this arises with binary exposure and disease, along with internal validation. In the unconstrained case of differential misclassification, there are closed-form expressions for maximum likelihood estimators of all parameters (Lyles 2002). This arises because the model for $(X, W|Y)$ is *saturated*, with six unknown parameters in total in the binary case (parameterized, for instance, as exposure prevalence, sensitivity, and specificity, all of which may vary with Y). Analogously, a parameterization in terms of positive and negative predictive values along with prevalence of apparent exposure leads directly to closed-form estimators, for again it produces a saturated model. The maximum likelihood (ML) estimators of the parameters for the $(W|Y)$ distribution are the sample analogues of the corresponding cell probabilities based on all observations, while those for $(X\$|W, Y)$ are the sample analogues from the validation observations. On the other hand, imposing the assumption of non-differential misclassification reduces the number of unknown parameters to four, and the loss of saturation implies that numerical optimization is required to obtain ML estimates.

One way to obtain ML estimates when validation data are available is via algorithms for missing-data problems (Carroll et al. 2006). Very roughly put, such algorithms iterate back and forth between filling in (imputing) the missing data given the current estimated parameter values and updating the parameter estimates given the current filled-in (completed) data. For instance, with a random internal validation sample, the complete data would comprise (W, X, Y) observations for all subjects, whereas the observed data arise with X missing completely at random for some (usually most) subjects. The EM algorithm or any other optimization procedure can then be applied to the missing-data likelihood function (Little and Rubin 2002), as can more *ad hoc* but popular techniques like multiple imputation (Cole et al. 2006). Bayesian computation can be implemented via similar strategies, with Markov chain Monte Carlo (MCMC) sampling of all unobservables given all observables (see Carlin and Louis (2008) for a general introduction to MCMC, and Gustafson (2003) for application of MCMC to misclassification models). This devolves to iteratively generating Monte Carlo samples of complete data given parameters and incomplete data, and parameters given complete data.

There are a variety of computational approaches to problems in which a prior distribution is used to replace point constraints such as non-differentiality. While MCMC provides a theoretical guarantee of providing a fully Bayesian answer given sufficiently long computing time, reliable implementation can be difficult and it is not very accessible to non-experts. The approximate approach of Monte Carlo sensitivity analysis (Greenland 2003, 2005; Greenland and Lash 2008; Lash et al. 2009) is simpler and intuitive, and works well for many problems, although again it may suffer from range violations of the sort referred to in Sect. 17.3.1, i.e., combinations of sample W prevalences and misclassification probabilities can arise which imply X prevalences below zero or above one. An alternative that avoids these issues is to perform an approximate Bayesian analysis by representing the prior distribution as extra data points and then applying standard likelihood techniques to the augmented set of data (Greenland 2009a, b).

Setting aside the issue of implementation, many contexts involving misclassified variables do seem to require a Bayes or Bayes-like approach since the available information about the relationship between ascertained and actual exposure is amorphous and can only be captured via the specification of a prior distribution. We have already seen a stylized example of this form in Sect. 17.3.1, where the sensitivity and specificity of exposure classification were not known exactly, but were not completely unknown: They could be bounded probabilistically, i.e., plausible prior distributions could be specified. For instance, we could model suppositions such as "sensitivity is likely between 75% and 90%." This theme is explored in Greenland (2009a, b), who considers prior specification for classification parameters at length. An important feature of this work is that, if so desired, prior distributions need be specified explicitly only for those parameters that will not be strongly informed by the data; the remaining parameters are in effect left with improper ("non-informative") priors.

A different situation where Bayesian methods can be used to reflect amorphous assumptions arises in the context of validation data. Often, the assumption of non-differential misclassification seems unjustifiable, but at the same time large departures from non-differentiality are not thought plausible. Chu et al. (2010) investigate prior distributions which reflect this situation and consequently produce an analysis which compromises between those arising from the non-differential assumption and those arising when this assumption is omitted (i.e., allowing arbitrary differentiality). For instance, the joint prior distribution for sensitivity among controls and sensitivity among cases could be assigned such that the two quantities have the same distribution marginally, while also having a high correlation between them. This implies an *a priori* assumption that the two sensitivities are unlikely to be very far apart, with the details of this assertion being controlled by the specific choice of correlation in the prior distribution. The same style of prior can be applied to the two specificities.

17.4 Other Topics

17.4.1 Misclassification of Outcomes

In epidemiological contexts, questions of exposure classification are often paramount. In some settings, however, the outcome variable Y may be subject to misclassification. For instance, Y might be assessed via an imperfect diagnostic test. This leads to an interesting distinction. If Y is a continuous variable, then classical measurement error models for Y lead to no bias in estimators based on a surrogate U for Y. To see this, suppose $U = Y + \varepsilon_U$ where ε_U has zero mean and is independent of all variables including Y. Then, the regression structure is unchanged, i.e., the mean of U given X is the same as the mean of Y given X. In particular, if $Y = \mathrm{f}(X;\beta) + \varepsilon_\mathrm{Y}$, then $U = \mathrm{f}(X;\beta) + \varepsilon_Y + \varepsilon_U$, and so inferences on β from fitting $\mathrm{f}(X;\beta)$ to U are also inferences on the dependence of Y on X, albeit power and precision is reduced by the presence of the additional error ε_U in U. In contrast, the regression of Y or U on an unbiased surrogate $W = X + \varepsilon_X$ for X would produce an attenuated (flattened) regression function (Carroll et al. 2006), and so we should not ignore the measurement error in X. Thus, there is a fundamental asymmetry of X and Y under the simplest continuous error models.

In contrast to the asymmetric nature of regression when X and Y are continuous, when X and Y are binary and their odds ratio describes their dependence, X and Y do play symmetric roles. This symmetry is in part because of the use of a symmetric parameter (the odds ratio). But the fact that independent non-differential misclassification of Y cannot be ignored reflects in part that such misclassification cannot in general be represented in the independent additive form of classical continuous regression.

Put more broadly, apart from some special cases, misclassification of Y changes the regression structure. As a simple way to convey the essence of the situation, say that U arises from Y via non-differential misclassification. Then, the regression of U on X is an affine transform of the regression of Y on X,

$$\Pr(U = 1|X = x) = (1 - Sp) + (Sn + Sp - 1)\Pr(Y = 1|X = x),$$

and hence cannot be linearized exactly if Sn and Sp are unknown (here, X may represent any number of multiple exposures with different data types). Thus, we see a general sense in which a given change in risk of the outcome $Y = 1$ with X is weakened by a factor of $(Sn+Sp-1)$ to yield the change in risk of apparent outcome

$U = 1$ with X. While this multiplicative factor applies to the risk itself, Neuhaus (1999) considers the impact on other scales, such as logit-risk.

The consequence is that, for discrete outcomes, correction for outcome misclassification can be as imperative as for exposure misclassification. It should be noted however that attempts to correct for outcome misclassification in case-control studies can lead to serious bias if applied naively, without taking account of the fact that cases are sampled in a much higher proportion than are controls, i.e., that Y affects the sampling probability (Greenland and Kleinbaum 1983). Furthermore, studies that collect their own outcome data (instead of relying on records) may apply much more stringent classification procedures than used in ordinary clinical practice, making use of external validation data especially hazardous.

17.4.2 Misclassification of Confounders

Poor measurement can be a serious issue for confounding variables as well as for exposure and/or outcome variables (Greenland 1980; Savitz and Baron 1989; Brenner 1993; Marshall et al. 1999; Fewell et al. 2007). As a general rule, adjusting for a misclassified version of a confounder will only partially control for the actual confounder. To illustrate, suppose that binary exposure X and binary outcome Y are not subject to misclassification, but the binary confounder Z is unobserved, with non-differentially misclassified surrogate V taking its place. In the present context, the non-differential assumption corresponds to conditional independence of V and (X, Y) given Z. Thus, the misclassification can be described by the sensitivity *Sn* and specificity *Sp* of V as a surrogate for Z.

For simplicity, suppose that the (X, Y) odds ratio is homogeneous across the Z strata, i.e., $OR(X, Y|Z = 0) = OR(X, Y|Z = 1)$, so that Z may be a confounder, but does not modify the odds ratio. Necessarily then, as *Sn* and *Sp* both increase to one, $OR(X, Y|V = \mathrm{v})$ must tend to $OR(X, Y|Z)$, for both $V = 1$ and $V = 0$. In the other direction, however, as *Sn* and *Sp* decrease to the scenario in which V is completely random noise, whereby the distribution of V no longer depends on Z (i.e., $Sn = 1 - Sp$), stratification by V corresponds to an entirely random division of the population. Thus, for both $V = 1$ and $V = 0$, $OR(X, Y|V = v)$ must tend to the crude (unadjusted) odds ratio relating X to Y. Intuitively then, between the extremes of random and perfect assessment of Z, we should see stratification by V producing results in between the crude $OR(X, Y)$ and the fully adjusted $OR(X, Y|Z)$.

With differential misclassification, this argument no longer holds and it is at least theoretically possible for a V-stratum specific odds ratio to fall entirely outside the range of the crude and fully adjusted odds ratios. The argument can also break down for polytomous variables even if non-differentiality holds (Brenner 1993), although the deviations from it do not appear large in plausible examples.

For the non-differential binary case, Fig. 17.3 displays $OR(X, Y|V = v)$ as *Sn* and *Sp* vary, for specific choices of $(Z|X)$ and $(Y|X, Z)$ distributions. A setting where confounder misclassification leads to a spurious appearance of a relationship between X and Y is given, as is a setting where confounder misclassification

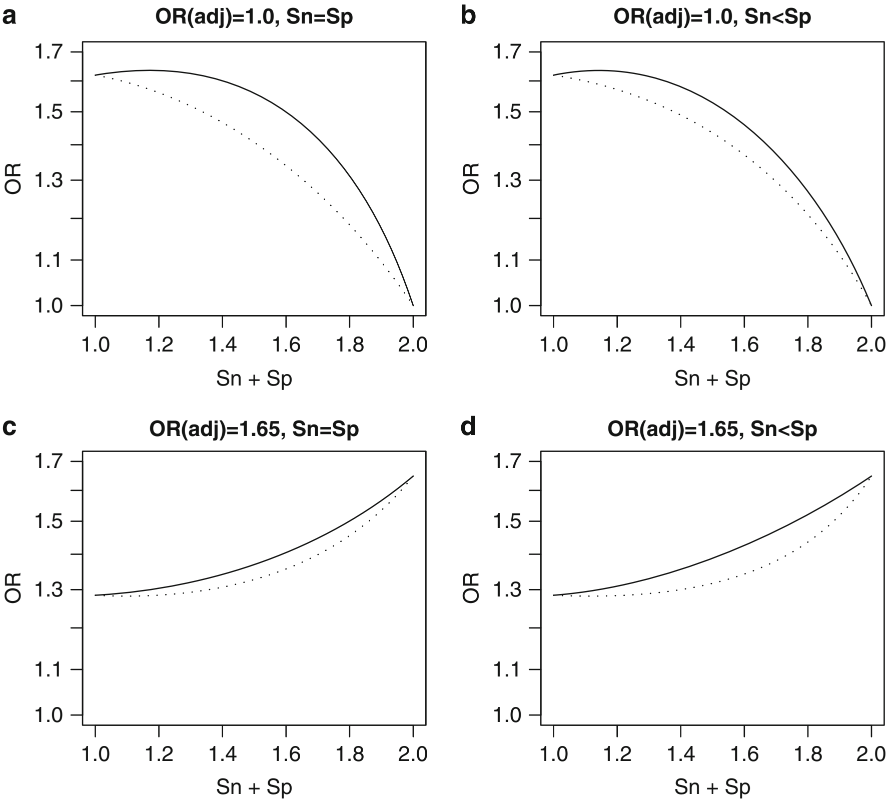


Fig. 17.3 Odds ratio for (X, Y) stratified by V, where V is a non-differentially misclassified surrogate for true confounder Z. Results are shown as the sensitivity and specificity of V vary, with the solid (*dotted*) curve corresponding to the $V = 0$ ($V = 1$) stratum. Note the use of logarithmic axis. In all cases, $\Pr(Z = 1|X = 0) = 0.25$ and $\Pr(Z = 1|X = 1) = 0.75$. In the upper panels, $\text{logitPr}(Y = 1|X, Z) = \text{logit}(0.1) + Z$, so X and Y are unassociated given Z. In the lower panels, $\text{logitPr}(Y = 1|X, Z) = \text{logit}(0.1) + 0.5X + 0.5Z$, i.e., X and Y have a homogeneous odds ratio of $\exp(0.5) \approx 1.65$ within levels of Z. In the *left-hand panels*, $Sn = Sp$. In the *right-hand panels*, Sp exceeds Sn, specifically $(Sn-0.4) = 1.5(Sp-0.6)$. In all situations, as the classification worsens to random ($Sn + Sp = 1$), $OR(X, Y|V = v)$ tends to the crude X-Y odds ratio. As the classification improves to perfect ($Sn + Sp = 2$), $OR(X, Y|V = v)$ tends to the fully adjusted $OR(X, Y|Z)$

attenuates the strength of the fully adjusted $(X, Y|Z)$ relationship. The figure also illustrates the fact that confounder misclassification can induce spurious heterogeneity (Greenland 1980). Starting with the homogeneous situation where $OR(X, Y|Z = z)$ is constant across Z, one sees that nonetheless $OR(X, Y|V = v)$ can depend on V even if the sensitivity and specificity for Z do not vary with X or Y. Analogous results apply to other measures of effect as well as odds ratios and show that confounder misclassification can also mask heterogeneity (Greenland 1980).

17.5 Conclusions

We have reviewed basic features and principles of misclassification, along with methods that attempt to account for it. As with its companion topic of measurement error, we would emphasize that typical adjustment methods are highly sensitive to underlying assumptions and models. In particular, simple assumptions such as non-differentiality, error independence, and error additivity often have no basis in fact and are easily violated, yet can have enormous impact on both the bias and precision of inferences. Thus, while an adjustment may often be better than no adjustment at all, we recommend sensitivity analysis, i.e., evaluation of how stable the results of adjustment are as the assumptions producing the adjustment are varied. Also, as illustrated in Sect. 17.3.1 and alluded to elsewhere, Bayesian techniques can be used to relax assumptions and thus better represent situations in which misclassification parameters are known approximately rather than exactly (such as when misclassification is thought to be close to, rather than exactly, non-differential). Such weakening of assumptions can (and we argue should) be made even if frequentist identification is sacrificed, for Bayesian identification can be maintained by using plausible priors in place of the questionable assumptions (Greenland 2009a, b; Gustafson 2009).

References

Birkett NJ (1992) Effect of non-differential misclassification on estimates of odds ratios with multiple levels of exposure. Am J Epidemiol 136:356–362

Brenner H (1993) Bias due to non-differential misclassification of polytomous confounders. J Clin Epidemiol 46:57–63

Brenner H (1996) Correcting for exposure misclassification using an alloyed gold standard. Epidemiology 7:406–410

Brenner H, Gefeller O (1993) Use of positive predictive value to correct for disease misclassification in epidemiologic studies. Am J Epidemiol 138:1007–1015

Brenner H, Savitz DA, Jöckel KH, Greenland S (1992) Effects of non-differential exposure misclassification in ecologic studies. Am J Epidemiol 135:85–95

Broemeling LD (2007) Bayesian biostatistics and diagnostic medicine. Chapman and Hall/CRC, Boca Raton

Bross IDJ (1954) Misclassification in 2 × 2 tables. Biometrics 10:478–486

Carlin BP, Louis TA (2008) Bayesian methods for data analysis, 3rd edn. Chapman and Hall/CRC, Boca Raton

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu C (2006) Measurement error in nonlinear models, 2nd edn. Chapman and Hall/CRC, Boca Raton

Chavance M, Dellatolas G, Lellouch J (1992) Correlated non-differential misclassification of disease and exposure. Int J Epidemiol 21:537–546

Chu H, Cole SR, Wei Y, Ibrahim JG (2009) Estimation and inference for case-control studies with multiple non-gold standard exposure assessments: with an occupational health application. Biostatistics 10:591–602

Chu R, Gustafson P, Le N (2010) Bayesian adjustment for exposure misclassification in case-control studies. Stat Med 29:994–1003

Cole SR, Chu H, Greenland S (2006) Multiple-imputation for measurement error correction (with comment). Int J Epidemiol 35:1074–1082

Cook J, Stefanski LA (1995) A simulation extrapolation method for parametric measurement error models. J Am Stat Assoc 89:1314–1328

Dendukuri N, Joseph L (2001) Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics 57:158–167

Dosemeci M, Wacholder S, Lubin JH (1990) Does non-differential misclassification of exposure always bias a true effect toward the null value? Am J Epidemiol 132:746–748

Drews C, Greenland S (1990) The impact of differential recall on the results of case-control studies. Int J Epidemiol 19:1107–1112

Fewell Z, Davey Smith G, Sterne J (2007) The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. Am J Epidemiol 166:646–655

Flegal KM, Keyl PM, Nieto FJ (1991) Differential misclassification arising from non-differential errors in exposure measurement. Am J Epidemiol 134:1233–1244

Greenland S (1980) The effect of misclassification in the presence of covariates. Am J Epidemiol 112:564–569

Greenland S (1982) The effect of misclassification in matched-pair case-control studies. Am J Epidemiol 116:402–406

Greenland S (1988) Statistical uncertainty due to misclassification: implications for validation substudies. J Clin Epidemiol 41:1167–1176

Greenland S (2001) Sensitivity analysis, Monte Carlo risk analysis, and Bayesian uncertainty assessment. Risk Anal 21:579–583

Greenland S (2003) The impact of prior distributions for uncontrolled confounding and response bias: a case study of the relation of wire codes and magnetic fields to childhood leukemia. J Am Stat Assoc 97:47–54

Greenland S (2005) Multiple bias modeling for analysis of observational data (with discussion). J R Stat Soc Ser A 168:267–308

Greenland S (2008) Maximum-likelihood and closed-form estimators of epidemiologic measures under misclassification. J Stat Plan Inference 138:528–538

Greenland S (2009a) Bayesian perspectives for epidemiologic research. III. Bias analysis via missing-data methods. Int J Epidemiol 38:1662–1673. doi: 10.1093/ije/dyp278

Greenland S (2009b) Relaxation priors and penalties for plausible modeling of nonidentified bias sources. Stat Sci 24:195–210

Greenland S, Gustafson P (2006) Accounting for independent non-differential misclassification does not increase certainty than an observed association is in the correct direction. Am J Epidemiol 164:63–68

Greenland S, Kleinbaum DG (1983) Correcting for misclassification in two-way tables and matched-pair studies. Int J Epidemiol 12:93–97

Greenland S, Lash TL (2008) Bias analysis. Chapter 19. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 3rd edn. Lippincott-Wolters-Kluwer, Philadelphia, pp 345–380

Gustafson P (2003) Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments. Chapman and Hall/CRC, Boca Raton

Gustafson P (2009) What are the limits of posterior distributions arising from nonidentified models, and why should we care? J Am Stat Assoc 104:1682–1695

Gustafson P, Greenland S (2006) Curious phenomena in adjusting for exposure misclassification. Stat Med 25:87–103

Gustafson P, Le ND (2002) Comparing the effects of continuous and discrete covariate measurement error with emphasis on dichotomization of mismeasured predictors. Biometrics 28:878–887

Gustafson P, Le ND, Saskin R (2001) Case-control analysis with partial knowledge of exposure misclassification probabilities. Biometrics 57:598–609

Hanson TE, Johnson WO, Gardner IA, Georgiadis MP (2003) Determining the infection status of a herd. J Agric Biol Environ Stat 8:469–485

Hui SL, Walter SD (1980) Estimating the error rates of diagnostic tests. Biometrics 36:167–171

Jones G, Johnson WO, Hanson TE, Christensen R (2010) Identifiability of models for multiple diagnostic testing in the absence of a gold standard. Biometrics 66:855–863

Kraus JF, Greenland S, Bulterys M (1989) Risk factors for sudden infant death syndrome in the U.S. collaborative perinatal project. Int J Epidemiol 18:113–120

Kristensen P (1992) Bias from non-differential but dependent misclassification of exposure and outcome. Epidemiology 3:210–215

Küchenhoff H, Mwalili SM, Lesaffre E (2006) A general method for dealing with misclassification in regression: the misclassification SIMEX. Biometrics 62:85–96

Lash TL, Fox MP, Fink AK (2009) Applying quantitative bias analysis to epidemiologic data. Springer, New York

Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York

Lyles RH (2002) A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure. Biometrics 58:1034–1037

Marshall JR (1990) Validation study methods for estimating exposure proportions and odds ratios with misclassified data. J Clin Epidemiol 43:941–947

Marshall JR, Hastrup JL, Ross JS (1999) Mismeasurement and the resonance of strong confounders: correlated errors. Am J Epidemiol 150:88–96

Natarajan L (2009) Regression calibration for dichotomized mismeasured predictors. Int J Biostat 5(1):Article 12

Neuhaus JM (1999) Bias and efficiency loss due to misclassified responses in binary regression. Biometrika 86:843–855

Newell DJ (1962) Errors in interpretation of errors in epidemiology. Am J Public Health 52: 1925–1928

Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. Oxford University Press, Oxford

Savitz DA, Baron AE (1989) Estimating and correcting for confounder misclassification. Am J Epidemiol 129:1062–1071

Tu X, Litvak E, Pagano M (1994) Studies of AIDS and HIV surveillance screening tests: can we get more by doing less? Stat Med 13:1905–1919

Tu X, Litvak E, Pagano M (1995) On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application in HIV screening. Biometrika 82:287–297

Wacholder S, Armstrong B, Hartge P (1993) Validation studies using an alloyed gold standard. Am J Epidemiol 137:1251–1258

Wacholder S, Dosemeci M, Lubin JH (1991) Blind assignment of exposure does not prevent differential misclassification. Am J Epidemiol 134:433–437

Walker AM, Blettner M (1985) Comparing imperfect measures of exposure. Am J Epidemiol 121:783–790

Weinberg CR, Umbach DM, Greenland S (1994) When will non-differential misclassification of an exposure preserve the direction of a trend? (with discussion). Am J Epidemiol 140:565–571

Zhou XH, Obuchowski NA, McClish DK (2002) Statistical methods in diagnostic medicine. Wiley, New York

Confounding and Interaction 18

Neil Pearce and Sander Greenland

Contents

N. Pearce (✉)
Department of Medical Statistics, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

Centre for Public Health Research, School of Public Health, Massey University Wellington, Wellington, New Zealand

S. Greenland
Department of Epidemiology, School of Public Health, University of California, Los Angeles, CA, USA

Department of Statistics, UCLA College of Letters and Science, Los Angeles, CA, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_10,

18.1 Introduction

All epidemiological studies are (or should be) based on a particular *source population* followed over a particular *risk period*. The goal is usually to estimate the effect of one or more exposures on one or more health outcomes. When we are estimating the effect of a specific exposure on a specific health outcome, *confounding* can be thought of as a mixing of the effects of the exposure being studied with the effect(s) of other factor(s) on the risk of the health outcome of interest. *Interaction* can be thought of as a modification, by other factors, of the effects of the exposure being studied on the health outcome of interest and can be subclassified into two major concepts: *biological interdependence of effects*, which includes concepts of synergism and antagonism, and *effect-measure modification*, also known as heterogeneity of a measure across the levels of another factor. Both confounding and interaction can be assessed by stratification on these other factors (i.e., the potential confounders or effect-measure modifiers). This chapter covers the basic concepts of confounding and interaction and provides a brief overview of analytical approaches to these phenomena. Because these concepts and methods involve far more topics than we can cover in detail, we provide many references to further discussion beyond that in the present handbook, especially to relevant chapters in Modern Epidemiology by Rothman et al. (2008).

18.2 Confounding

18.2.1 Basic Concepts

When estimating the effect of an exposure on those exposed, confounding occurs when the exposed and non-exposed subpopulations of the source population have different background disease risks, which is to say these subpopulations would have different disease risks even if exposure effects had been absent from both subpopulations (Greenland and Robins 1986; Greenland et al. 1999a, b; Maldonado and Greenland 2002; Rothman et al. 2008, Chap. 4). When we estimate the effect of exposure on the exposed by comparing the frequency of disease in the exposed and non-exposed groups, we assume that the disease frequency in the non-exposed group provides a valid estimate of what the disease frequency would have been in the exposed group if it had not been exposed. If this assumption is incorrect, that is, if the exposed and non-exposed groups would have had different disease frequencies in the counterfactual situation in which the exposed group had not been exposed, then we say that the comparison of the exposed group to the non-exposed group is confounded.

More generally, confounding can arise when the exposed and non-exposed groups are not completely comparable or "exchangeable" with respect to their response to exposure; that is, for at least one level of exposure, the exposed and unexposed groups would exhibit different risks even if they both had experienced that exposure level (Greenland et al. 1999b; Maldonado and Greenland 2002; Greenland and Robins 2009). Note that the earlier definition takes this level to be that of non-exposure because it is presumed there that the effect of exposure on the exposed is the effect of interest. If instead we were interested in the effect of non-exposure on the non-exposed (as might be the case in a study of a preventive factor), we would have confounding if the exposed group failed to exhibit the risk that the non-exposed would have had if they had been exposed.

This problem of non-comparability (non-exchangeability) can also occur in randomized trials because randomization may fail, leaving the treatment groups with different characteristics (and different baseline disease risk) at the time that they enter the study, or because of differential loss and non-compliance across treatment groups. However, there is more concern about non-comparability in observational epidemiology studies because of the absence of randomization. Randomization prevents certain sources of confounding (e.g., confounding due to physician selection of treatment based on patient characteristics, also known as "confounding by indication"); also, bias due to differential loss and non-compliance can be at least partially controlled by using the randomization indicator (the "intent-to-treat" variable) as an instrumental variable (Sommer and Zeger 1991; Greenland 2000a). These benefits of randomization are not available in observational studies, and in fact confounding should be expected to occur as a by-product of ordinary life events and choices.

As an example, if we compare the risk of lung cancer in people with a low dietary beta carotene intake compared with people with a high dietary beta carotene intake, it is very likely that these two groups will differ with respect to other risk factors for lung cancer such as tobacco smoking because people who are less health conscious are more likely to smoke as well as to neglect dietary recommendations. If this is the case (e.g., if a greater percentage of people smoke in the low beta carotene intake group than in the high beta carotene intake group), then smoking will confound the association between beta carotene intake and lung cancer: the higher smoking prevalence among those with a low beta carotene diet will lead to a higher lung cancer risk among them compared to those with a high beta carotene diet, even if beta carotene intake itself has no effect on lung cancer risk. This in turn will create a spurious inverse association between beta carotene and lung cancer if we do not or cannot fully adjust for smoking differences.

Any variable that affects disease in the absence of exposure has the potential to confound the exposure-disease relationship. It will confound that relationship among the exposed if, in the absence of exposure, it would have a distribution that is sufficiently different across exposure groups to produce a difference in risk across those groups even if exposure were absent. This was the case for smoking in the beta carotene example. A confounder, if not adequately controlled, will bias the estimated effect of exposure on disease. The bias will be upward if the higher-risk levels of the confounder occur more frequently among the exposed; conversely, the bias will be downward if the higher-risk levels of the confounder occur more frequently among the unexposed.

Confounding may even reverse the apparent direction of an effect in extreme situations. Confounding may also occur when the main exposure under study has no effect on the risk of disease – a spurious association may be observed which is entirely due to confounding. Factors associated with confounders can also act like confounders and serve as surrogates for confounders, provided that they are not affected by exposure or disease. For example, socioeconomic status may serve as a surrogate measure of causal factors (living conditions, lifestyle, lack of preventive care, etc.) that are potential confounders.

Three conditions are traditionally given as necessary (but not sufficient) for a factor to be a confounder of the exposure effect in the exposed (Rothman et al. 2008, Chap. 9). First, to produce confounding, a factor has to be predictive of disease in the absence of the exposure under study. Note that a confounder need not be a genuine cause of the disease under study, but merely "predictive" within exposure levels apart from chance relations. Hence, surrogates for causal factors (e.g., ethnicity, gender, socioeconomic status) may be regarded as potential confounders, even if they are not direct causal factors. It is not always clear from the data whether an observed relation between the factor and disease represents a genuine (replicable) predictive quality, as opposed to (say) chance. In making such a determination, prior information as opposed to statistical testing should play a dominant role (Greenland and Neutra 1980; Miettinen and Cook 1981; Robins and Morgenstern 1987). This is why one almost always sees adjustment made for age and sex: these factors are known to be predictive of risk of most diseases. When prior information is not available, one must of course turn to the data collected for the study as a guide as to whether the factor is predictive of disease in the source population; even in these cases, however, there are better strategies for confounder selection than those based on statistical testing. We will return to this topic below.

Second, a confounder has to be associated with the study exposure in the source population. It may occur that when participants in a case-control study are selected from the source population, then due to chance, a factor may be associated with exposure in the study, even though it was not associated with exposure in the source population. In this situation, the factor is not a confounder (Miettinen and Cook 1981; Robins and Morgenstern 1987). Although in practice it is common to use the data actually collected to decide whether a factor is associated with exposure, more commonly the data are used to decide whether adjustment for the factor makes an important difference in the estimated exposure effect, a practice we will discuss below.

In a case-control study, one should expect a confounder to be associated with exposure among the controls (at least if the controls are selected with no bias). If the factor is not associated with exposure among controls, an association may still occur among the cases simply because the study factor and a potential confounder are both risk factors for the disease, but this is a consequence of those effects and so does not cause confounding. A factor-exposure association will only indicate confounding by the factor if it reflects the association in the source population.

Third, a variable that is affected by the exposure or by the disease, for example, an intermediate in the causal pathway between exposure and disease, or conditions that are caused by the health outcome of interest, should not be treated as a confounder because to do so could introduce serious bias into the results (Greenland and Neutra 1980; Robins and Morgenstern 1987; Robins and Greenland 1992; Weinberg 1993; Cole and Hernan 2002; Rothman et al. 2008, Chap. 9). For example, in a study of obesity and death from coronary heart disease, it would be inappropriate to control for hypertension if hypertension was a consequence of obesity and, hence, a part of the causal chain leading from obesity to death from coronary heart disease. On the other hand, if hypertension itself was of primary interest, then it would be studied directly; obesity would be regarded as a potential confounder if it also involved exposure to other risk factors for death from coronary heart disease.

Similarly, we should avoid controlling for health outcomes that may be part of disease genesis, such as reduced pulmonary function following exposure to a respiratory hazard in a study of chronic obstructive lung disease (Checkoway et al. 2004). We would, however, be justified in controlling for baseline (i.e., pre-exposure) lung function if there were reasons to believe that baseline lung function was associated with subsequent exposure level. Evaluating whether certain factors are exposure or health outcome intermediates in causal pathways requires information external to the study. Intermediate variables can sometimes be included in the analysis, although special techniques are then required to avoid adding bias (Robins 1989; Robins and Greenland 1992, 1994; Robins et al. 2000). In no case would control of a variable affected by the disease be valid, however (Greenland et al. 1999a; Pearl 2009).

Assessment of confounding by a factor that is not an intermediate involves consideration of whether the exposed and non-exposed groups are "comparable" in the source population with respect to their disease risk in the absence of exposure. In practice, we often focus on specific potential confounders – variables that are risk predictive of disease in the absence of exposure (such as age and sex) and assess whether they are associated with exposure in the source population on which the study was based. If such an association is present, it is evident that the two groups are not comparable or exchangeable with respect to baseline risk. If such an association is absent, however, it does not mean that the groups are comparable because there may be other uncontrolled risk factors that confound the observed association, or the association may have been obscured by measurement error.

Because it involves judgments about causal as well as temporal ordering, the property of being a confounder cannot be determined from data alone (Greenland and Neutra 1980; Miettinen and Cook 1981; Greenland and Robins 1986; Greenland et al. 1999a; Robins 2001; Hernan et al. 2002; Pearl 2009). Once that ordering is established, however, it is common to assess confounding by seeing whether the main effect estimate changes when the potential confounder is controlled in the analysis. In this approach, near equality of the crude and adjusted effect estimates is taken as evidence that there is no confounding by the factor; conversely, an important difference is taken as evidence of confounding by the factor. Many epidemiologists prefer to make a decision based on the basis of this "collapsibility" or "change-in-estimate" criterion (rather than the criterion of "exchangeability"), although this approach can be misleading, particularly if (as usual) there is misclassification of the adjustment factors or the exposure (Greenland 1980; Greenland and Robins 1985; Savitz and Baron 1989; Marshall and Hastrup 1996, 1999) or if the outcome is common and the measure is an odds ratio or rate ratio (Miettinen and Cook 1981; Greenland and Robins 1986; Greenland 1996; Greenland et al. 1999b); also, this criterion does not exhibit good statistical properties, although it is no worse than significance-testing procedures (Maldonado and Greenland 1993).

The decision to control for a presumed confounder can be made with more confidence if there is supporting prior knowledge that the factor is predictive of disease, independent of its association with exposure. Such prior knowledge is usually available for well-studied factors such as age, sex, and tobacco smoking. At the very least, it is usually known if the factor cannot have been affected by exposure, for example, if the factor represents an event occurring before exposure. If even this much is uncertain, the decision to control or not control a variable may be controversial, in which case analyses both with and without its control may be presented (Greenland and Neutra 1980).

As a final caution, in studies involving aggregate-level effects (such ecologic and multilevel studies), a factor at one level may, if not controlled, confound effect estimates at another level, and a factor may modify and confound effects differently at different levels of aggregation. For example, both the income of an individual and the income of his or her neighborhood may separately predict risk of an outcome, possibly in opposite directions. Robbery rates are often higher in low-income neighborhoods, yet within neighborhoods it could still be that an individual's risk of robbery went up as his or her income went up. In that case, both neighborhood income and individual income could be confounders, but would confound effect estimates in opposite directions if both were positively and separately associated with the exposure under study. Thus, regardless of level of interest (e.g., country, neighborhood, individual), it is often essential to measure and adjust for variables on other levels (Greenland 2001).

18.2.2 Example of Confounding

Table 18.1 presents a hypothetical example of confounding in a cross-sectional study of asthma. Overall, one-half of the study participants are smokers and one-half are not. However, two-thirds of those in the exposed group are smokers compared with one-third of the non-exposed workers. Thus, although exposure is not associated with asthma either among smokers (the prevalence of asthma is 40% in the exposed and 40% in the non-exposed, prevalence ratio (PR) = 1.0) or in non-smokers (the prevalence of asthma is 20% in the exposed and 20% in the non-exposed, PR = 1.0), it is associated with asthma (PR = 1.25) when the two subgroups are combined. This occurs because smoking is associated with the exposure in the source population and is an independent risk factor for asthma. In this hypothetical example, each stratum-specific estimate is 1.00; thus the adjusted estimate will also be 1.00 whatever weights are used. Thus, the crude

Table 18.1 Hypothetical example of confounding by tobacco smoking in a study of occupational asthma

	Smokers		Non-smokers		Total	
	Exposed	Non-exposed	Exposed	Non-exposed	Exposed	Non-exposed
Asthma cases	800	400	200	400	1,000	800
Non-cases	1,200	600	800	1,600	2,000	2,200
Total	2,000	1,000	1,000	2,000	3,000	3,000
Prevalence (%)	40	40	20	20	33.3	26.7
Prevalence ratio	1.0		1.0		1.25	

prevalence ratio is 1.25, whereas the adjusted prevalence ratio is 1.00, indicating that confounding has occurred (provided that there has not been biased selection of the study participants from the source population).

18.2.3 Control in the Study Design

Confounding can be controlled in the study design, in the analysis, or both. There are three common methods for control at the design stage (Rothman et al. 2008, Chap. 11). The first is randomization – random allocation of participants to exposure categories. However, this is usually only an option for potentially beneficial exposures, for example, it would be impractical and unethical to conduct a randomized trial of the health effects of smoking, and as mentioned earlier, randomization may fail to prevent all confounding.

A second method of control at the design stage is to restrict the study to narrow ranges of values of the potential confounders, for example, by restricting the study to white males aged 35–54. This approach has a number of conceptual and computational advantages, but may severely restrict the number of potential study subjects and ultimately limit the generalizability of the study.

A third method of control involves matching study subjects on potential confounders. For example, in a cohort study, one could match a white male non-exposed subject aged 35–39 with an exposed white male aged 35–39. This will prevent age-gender-ethnicity confounding in a cohort study, but is seldom done because it is expensive and time-consuming. In case-control studies, matching does not prevent confounding but does facilitate its control in the analysis in that matching on a strong confounder will usually increase the precision of effect estimates. However, matching may reduce precision in a case-control study if it is done on a factor which is associated with exposure, but is only a weak risk factor for the disease of interest.

If a matching factor is not a risk factor, the matching process effectively turns it into a confounder that must be controlled in the analysis, thus reducing precision and increasing analytical complexity. For example, in a case-control study of power-frequency electromagnetic field (EMF) exposure and childhood cancer, choosing sibling controls (i.e., for each case choosing a sibling as a control) would mean that in almost every instance, the case and control would have lived in the same house and would have similar EMF exposure, resulting in almost no exposure-discordant pairs and almost no precision in the resulting matched-pair estimates.

As already mentioned, matching may be expensive and time-consuming. Finding suitable controls becomes increasingly difficult as the number of matching factors increases beyond two or three. Moreover, when it occurs, the increase in precision from matching is often modest, typically involving a 5–15% reduction in the variance of the effect estimate (Schlesselman 1982; Thomas and Greenland 1983). Therefore, although there are many discussions of matching stress issues of statistical efficiency, practical considerations (such as ease of finding controls) are often more important (Rothman et al. 2008, Chap. 11).

18.2.4 Control in the Analysis

Confounding can also be controlled in the analysis by adjusting simultaneously for all confounding factors or a sufficient subset of them. This presumes, of course, that a sufficient subset has been accurately measured, which is often not the case. Methods for controlling confounding in the analysis are discussed in more depth in the chapters on specific study designs (chapters ▶Cohort Studies, ▶Case-Control Studies, and ▶Intervention Trials of this handbook), in Part II (Methodological Approaches in Epidemiology) of this handbook, and in many textbooks (e.g., Rothman et al. 2008, Chaps. 15–21). In the simplest situation, control of confounding in the analysis involves stratifying the data according to the levels of the confounder(s) and calculating an effect estimate that summarizes the association across strata of the confounder(s). As an example, controlling for age (grouped into five categories) and gender (with two categories) might involve grouping the data into the 10 $(= 5 \times 2)$ confounder strata and calculating a summary effect estimate, which is a weighted average of the stratum-specific effect estimates. It is usually not possible to control simultaneously for more than two or three confounders in a stratified analysis, since finer stratification will often lead to many strata containing no exposed or no non-exposed persons. Such strata are uninformative; thus, too fine stratification is wasteful of information. This problem can be mitigated to some extent by the use of regression modeling (cf. chapter ▶Regression Methods for Epidemiological Analysis of this handbook), which allows for simultaneous control of more confounders by "smoothing" the data across confounder strata.

18.2.5 Assessment of Confounding

When one lacks data on a suspected confounder, and thus cannot control confounding directly, it is still desirable to assess the likely direction and magnitude of the confounding. In particular, it may be possible to obtain information on a surrogate for the confounder of interest. For example, social class is associated with many lifestyle factors such as smoking and may therefore be a useful surrogate for some lifestyle-related confounders. Even though confounder control will be imperfect in this situation, it is still possible to examine whether the exposure effect estimate changes when the surrogate is controlled in the analysis and to assess the strength and direction of the change.

As an example, suppose the relative risk relating low dietary beta carotene intake to lung cancer actually increases (e.g., from 2.0 to 2.3) or remains stable (e.g., at 2.0) when social class is controlled. This might be taken as evidence that the observed excess risk is not entirely due to smoking because social class is associated with smoking (Kogevinas et al. 1997), and control for social class involves partial control for smoking. The strength of this evidence depends of course on what exposure is being studied and what sort of classification errors or other sources of bias are present.

Even if it is not possible to obtain confounder information for any study participants, it may still be possible to estimate the likely strength of confounding by particular risk factors. This is often done in occupational studies, where tobacco smoking is a potential confounder, but smoking information is rarely available. In fact, although smoking is the strongest risk factor for lung cancer, with relative risks of tenfold or more, it appears that smoking rarely exerts a confounding effect of greater than 1.5 times in studies of occupational disease (Axelson 1978, 1989; Siemiatycki et al. 1988; Kriebel et al. 2004) (although this degree of confounding may still be important in some contexts).

There are several approaches to the assessment of potential confounding by factors such as cigarette smoking when data are lacking or incomplete. One approach is to conduct an analysis of smoking-related diseases other than the disease of primary interest (Steenland et al. 1984). If mortality from such diseases (e.g., non-malignant respiratory disease) is not elevated, this may suggest that any excess for the disease of interest is unlikely to be due to smoking. Similarly, one might be less inclined to attribute an excess of an alcohol-related cancer to unusually high drinking prevalence among the exposed if liver cirrhosis mortality is not elevated.

When detailed individual risk factor information is not available on a potential confounder, it may be possible to assess the impact of this factor on risk estimates by conducting a type of *sensitivity analysis* that estimates the potential direction and extent of confounding (Cornfield et al. 1959; Bross 1967; Axelson 1978, 1989; Schlesselman 1978; Checkoway and Waldman 1985; Axelson and Steenland 1988; Flanders and Khoury 1990; Rothman et al. 2008, Chap. 19; Lash et al. 2009; see also chapter ▶Sensitivity Analysis and Bias Analysis of this handbook). In this sensitivity analysis, the magnitude of the effect of the potential confounder on the disease should be known with some confidence, and the prevalence of the potential confounder among the exposed and comparison groups should be estimable, within reasonable bounds. Then, a range of confounding effects, including a "worst case scenario," can be calculated (Checkoway et al. 2004).

The overall incidence rate, I, of disease in a population can be written as the sum of incidence rates at the different levels of the confounder, weighted by the proportion of the population at each confounder level (Axelson 1978). For a confounder which has two levels (present, absent) whose presence multiplies the rate by the relative risk due to the confounder (RR_c), we have

$$I = I_0(1 - p_c) + RR_c(I_0)(p_c)$$

where:

I = incidence rate overall
I_0 = incidence rate among those who are confounder negative
p_c = prevalence of the confounder in the source population from which the cases arose

This expression can be expanded to include several levels of a confounder, for example, light, moderate, and heavy smoking. By applying the equation to two (or more) study groups (e.g., exposed and non-exposed subgroups) for which $\mathrm{p_c}$ is assumed to differ, one can calculate a confounding bias factor B_c for comparisons of these two exposed groups. If there is no effect of exposure, then B_c is the magnitude of effect one would observe due to differences in $\mathrm{p_c}$ alone. Then, if RR_{Obs} is the observed relative risk comparing exposed to non-exposed,

$RR_{Adj} = RR_{Obs}/B_c$ is the adjusted relative risk, controlling for confounding.

To illustrate, suppose that we observe $RR_{Obs} = 3.2$ for lung cancer and some dichotomous occupational exposure, and we are concerned about how much of this elevation is due to confounding by smoking, but there are no smoking measurements in the study. We might assume that smoking habits in the non-exposed approximate those of other typical blue-collar workers whose smoking habits have been studied. To estimate the most extreme confounding that might reasonably be expected, we might assume that the exposed were heavier smokers, with habits more like the 90th percentile of blue-collar workers. These assumptions would imply that the non-exposed might have been 50% non-smokers, 40% moderate smokers, and 10% heavy smokers, whereas the exposed were 20% non-smokers, 55% moderate smokers, and 25% heavy smokers. Assuming that moderate smoking confers a relative risk of smoking of 10 compared to non-smokers, and heavy smoking a relative risk of 20, the confounding due to smoking is $B_c = 1.65$. Thus, one would observe a relative risk of 1.65 comparing exposed to unexposed due to these smoking differences alone. One can then calculate that $RR_{Adj} = 3.2/1.65 = 1.9$ as a hypothetically "adjusted" exposure effect under a plausible, but unlikely, scenario for smoking differences among exposed groups. If RR_{Adj} is elevated, one might conclude that confounding is unlikely to be the entire explanation for the elevated risk (Checkoway et al. 2004).

This method allows one to place limits on the degree of confounding that can result from failure to adjust for an unmeasured risk factor that is associated with the exposure under study. Its application is restricted however to control for factors whose risks are well established quantitatively and for which the confounder prevalence in the population can be estimated fairly reliably. The relative risk that would result from differences in the prevalence of a covariate, such as smoking or alcohol consumption, may be quite limited, even in the absence of complete knowledge about the covariate (Cornfield et al. 1959; Bross 1967; Flanders and Khoury 1990). In particular, Flanders and Khoury (1990) showed that

$$1 \leq B_c \leq \min\{OR, RR_c, 1/p_c, RR_c/(q_c + RR_c * p_c), OR/(q_c + RR_c * p_c)\}$$

where B_c, RR_c, and p_c are as above, $q_c = 1 - p_c$, and OR is the odds ratio measuring the association between exposure and the covariate (both measured dichotomously in this example). Using this equation, rather severe limits can be placed on the range of possible values of B_c with information about p_c and RR_c alone, making no assumption about OR at all. For example, if $p_c = 0.5$, and RR_c for esophageal cancer from 1 pack/day smoking $= 5$, then $B_c = 1.7$. That is, given $RR_c = 5$, an observed RR_{obs} for an exposure effect is unlikely to be confounded by more than 1.7 times. With reasonable assumptions about the likely values of OR and RR_c, the association between exposure and the confounder, the maximum value for B_c could be lower or higher. This method can be extended for the situation in which there are multiple levels of exposure and multiple levels of covariates (Schlesselman 1978; Flanders and Khoury 1990) and to other measures of association such as odds ratios (Yanagawa 1984).

A more sophisticated version of sensitivity analysis places uncertainty distributions (priors) on the unknown quantities in the sensitivity formula, repeatedly draws values from these distributions and corrects the data or estimates based on the drawn values, adds in a random-error correction, and presents the resulting distribution of corrected estimates. Such Monte-Carlo sensitivity analysis has long been a staple of risk analysis and is now finding application in epidemiology, along with related Bayesian methods (e.g., see Greenland 2003, 2005; Steenland and Greenland 2004; Rothman et al. 2008, Chap. 19; Lash et al. 2009; see also chapter ▶Sensitivity Analysis and Bias Analysis of this handbook).

Sensitivity analysis can also be useful in certain situations in which confounder information has been collected, but the validity or precision of those data are weak. Sometimes smoking data are available on only a subset of the members of a cohort. One option is to conduct an analysis that controls for smoking directly on the subset with smoking data. However, the precision of this analysis and the generalizability of findings to the entire cohort may be questionable. Instead, one might apply the data relating the confounder to exposure among this subset in a sensitivity analysis for the entire cohort (Fingerhut et al. 1991). A third option is to adjust the entire cohort based on the data from the subset, using two-stage methods or missing-data methods (Rothman et al. 2008, Chap. 15).

18.2.6 Relationship Between Confounding and Other Biases

In this chapter confounding has been defined as non-comparability (non-exchangeability) of the exposed and non-exposed subgroups in the source population with respect to their risk of the disease outcome in the absence of exposure. Confounding is thus a property of the source population rather than of the specific group of study participants.

Selection bias involves biases arising from the procedures by which the study participants are selected (or select themselves) from the source population. Thus, selection bias is not an issue in a cohort study (or cross-sectional study) involving complete recruitment and follow-up because in this instance the study group is the entire source population. However, selection bias can occur if participation in the study or follow-up is incomplete. Selection bias is usually more of an issue in case-control studies in that it may occur if the case group does not include (or is not representative of) all cases in the source population or if the control group is not representative of the population at risk that the cases came from.

If control selection involves only sampling at random from the entire source population with complete cooperation, selection bias is a minor concern. More realistically, however, selection bias is likely if response rates differ according to exposure level and disease status. When controls are selected from among persons with other diseases, considerable care must be taken in specifying the diseases that form the control group. In particular, a specific disease may not correctly reflect the exposure pattern in the source population, especially if it is caused by the study exposure. One strategy is to include only diseases that are thought to be unrelated to the exposure(s) of interest (Rothman et al. 2008, Chap. 8), but this requirement may be difficult to satisfy in practice due to lack of adequate evidence for the absence of exposure effects (Axelson et al. 1982).

An alternative method is to select as controls a sample of all other diseases. This approach is reliable if one can be sure the factor under study does not markedly increase risk of numerous or the most common diseases. It has become common practice to exclude diseases known to be related to exposure from the pool of potential controls; however, even this restriction will not always eliminate bias (Pearce and Checkoway 1988).

Selection bias and confounding are not always clearly demarcated. In particular, non-response at baseline of a cohort can be viewed as a source of confounding, since it may produce associations of exposure with other risk factors in the study cohort and thus turn those factors into confounders (Greenland et al. 1999a). A similar phenomenon occurs in case-control studies when selection is affected by a factor that itself affects exposure. An example occurs when matching on a factor that is associated with exposure in the source population; even if the factor is not a risk factor for disease in the absence of exposure, matching may turn it into a confounder which must be controlled in the data analysis because matching will create the necessary factor-disease association if exposure affects disease (Rothman et al. 2008, Chap. 11). Unfortunately, as discussed earlier, if selection is affected by exposure and associated with case-control status (e.g., selection bias due to inappropriate selection of controls from persons with other diseases, or selection on factors affected by exposure), stratification on the selection-related factors will rarely produce valid estimates, and hence this type of selection bias should not be viewed as confounding.

Information bias is the result of misclassification of study participants with respect to disease or exposure status. Thus, the concept of information bias refers to those people actually included in the study, whereas selection bias refers to the selection of the study participants from the source population, and confounding generally refers to non-comparability of subgroups within the source population. There are many methods to adjust for misclassification, for example, (Copeland et al. 1977; Greenland and Kleinbaum 1983; Espeland and Hui 1987; Greenland 1988;

Armstrong et al. 1992; Thomas et al. 1993; Armstrong 1998; Carroll et al. 2006). These methods require estimates of sensitivity and specificity, predictive values, or the reliability of the measurement, based on prior information or validation data. Estimates based on prior information are often only best guesses that may not apply to the population under study. However, it is an informative exercise to conduct sensitivity analyses that explore the range of results that might have occurred under various scenarios (Rothman et al. 2008, Chap. 19), and again, Monte-Carlo sensitivity analyses and Bayesian analyses may be applied (Gustafson 2003; Fox et al. 2005; Greenland 2005, 2009a; Rothman et al. 2008, Chap. 19; Lash et al. 2009; see also chapter ▶Sensitivity Analysis and Bias Analysis of this handbook).

Some consider any bias that can be controlled in the analysis as confounding, but this definition is too general because any bias control requires background information for proper execution, and any bias can be controlled in the analysis given enough background information. Confounding is distinguished in that it represents a mixing or confusion of the effects of other factors with the effects of the study exposure, a concept that goes back at least as far as the writings of John Stuart Mill (Greenland et al. 1999b). Other biases are then categorized according to whether they arise from the selection of study subjects (selection bias), their classification (information bias), or misapplication of statistical methods.

Most observational studies suffer from more than one form of bias, and the effects of multiple biases may compound error. Perhaps the best-appreciated situation is when there is misclassification of a confounder, in which case attempts to control for the confounder will not fully control that confounder and may actually increase bias (Greenland 1980; Greenland and Robins 1985; Savitz and Baron 1989; Marshall and Hastrup 1996, 1999). When (as is often the case) multiple biases are present, many complex and counterintuitive phenomena can occur, and a clear picture of the net effects of bias will require analyses that account for these bias interactions (Greenland 2005; Rothman et al. 2008, Chap. 19; Lash et al. 2009).

18.3 Interaction

18.3.1 Basic Concepts

Concepts of *interaction* and *effect-measure modification*, also known as effect heterogeneity and effect variation, refer to conditions in which the effect of exposure on the outcome under study varies by some other factor. In other words, in order to know or estimate the effect of exposure on an outcome (such as a disease time, a disease risk, or a disease rate), we must first know whether or not another factor is present (or what the level of this other factor is). This concept can be subclassified into two major concepts: *biological interdependence of effects*, which includes synergism and antagonism, and *effect-measure modification*, also known as heterogeneity of an effect measure. With regard to the latter, all secondary risk factors modify either the rate (or risk) ratio or the rate (or risk) difference when an exposure effect is present, and uniformity over one measure implies non-uniformity over the other (Steenland and Thun 1986; Rothman et al. 2008, Chap. 4), for example, an apparent additive joint effect implies a departure from a multiplicative model.

A further source of ambiguity is that the term "effect modification" implies that one factor in some way biologically "modifies" the effect of the other factor, but this is not necessarily the case. For this reason, the terms "effect-measure modification" and "effect-measure variation" are more accurate terms and are logically equivalent to the definition of "interaction" used in most statistics books and computer programs (Rothman et al. 2008, Chap. 5).

The concepts of interaction and confounding are quite distinct. An effect-measure modifier may or may not be a confounder and a confounder may or may not be an effect-measure modifier (Miettinen 1974; Rothman et al. 2008, Chap. 5). For example, if we are comparing exposed and non-exposed subgroups of a population, and the percentage of people who smoke (and the intensity of smoking and the length of time that each person has been smoking) is the same in both groups, then smoking is not a confounder. However, the rate ratio for the exposure effect may still vary by smoking status, for example, the exposure may double the risk of disease in smokers but have no effect in non-smokers. In this situation, smoking would not be a confounder but would be an effect-measure modifier.

18.3.2 Example of Effect-Measure Modification

Table 18.2 presents a hypothetical example of effect-measure modification in a cross-sectional study of asthma. The overall findings are the same as for the study presented in Table 18.1, but the stratum-specific findings are different. Now there is no confounding by smoking because the percentage of smokers is the same in the exposed and non-exposed groups. However, there is effect-measure modification since the prevalence ratio (for the association of exposure with disease) is 1.5 in smokers and 1.0 in non-smokers. Thus, whereas the assessment of confounding involved the comparison of the crude and adjusted effect estimates, the assessment of effect-measure modification involves the comparison of the stratum-specific effect estimates with each other.

Table 18.2 Hypothetical example of effect modification by tobacco smoking in a study of asthma prevalence

	Smokers		Non-smokers		Total	
	Exposed	Non-exposed	Exposed	Non-exposed	Exposed	Non-exposed
Asthma cases	600	400	400	400	1,000	800
Non-cases	900	1,100	1,100	1,100	2,000	2,200
Total	1,500	1,500	1,500	1,500	3,000	3,000
Prevalence (%)	40	26.7	26.7	26.7	33.3	26.7
Prevalence ratio	1.5		1.0		1.25	

18.3.3 Concepts of Interaction

Although at first glance the assessment of interaction is relatively straightforward, there are considerable hidden complexities. Some of the analytical issues in studying effect-measure modification will be illustrated with data (Table 18.3) from a study by Selikoff et al. (1980) of lung cancer death rates per 100,000 person-years at risk in relation to exposure to cigarette smoke and asbestos (Steenland and Thun 1986). The rate difference due to asbestos exposure is 472 per 100,000 person-years in non-smokers and 736 per 100,000 person-years in smokers. Thus, the rate difference for the effect of asbestos exposure on lung cancer mortality is lower in non-smokers than in smokers. On the other hand, the rate ratio for the same effect is higher in non-smokers (asbestos rate ratio = 17.5) than in smokers (asbestos rate ratio = 4.7). Thus, both the rate difference and the rate ratio are subject to effect-measure modification in that the effect estimate depends on the presence or absence (or more generally, the level) of another factor (i.e., smoking), but the dependencies are in opposite directions: the rate difference is larger in smokers and the rate ratio is larger for non-smokers.

For the most part, statisticians have taken this dependence of modification on the underlying effect measure to imply that the assessment of interaction is "model dependent" (Kupper and Hogan 1978; Walter and Holford 1978), and in most of statistics, the term "interaction" corresponds to effect-measure modification. In contrast to these statistically based definitions, other authors adopt a definition of *biological* interaction in which two factors are said to exhibit interdependent effects or "synergism" if their presence is required by the same sufficient cause and "antagonism" if the absence of one and the absence or presence of the other is required (Rothman 1976; Rothman et al. 2008, Chap. 2). With this conceptualization, one can show that the presence and degree of effects and effect-measure modification depend to a large extent on the prevalence of causal cofactors of exposure, as well as the actual biological mechanisms at work (Rothman et al. 2008; see chapter ▶Basic Concepts of this handbook for further explanation).

A closely related conceptualization defines biological interaction to be present when individual patterns of response (the potential or counterfactual outcomes) to one factor changes when the other "interacting" factor is changed (Greenland and Poole 1988; Rothman et al. 2008, Chap. 5). Under this conceptualization, non-additivity of risks will imply the presence of biological interaction, and under

Table 18.3 Example of joint effects: lung cancer mortality rates per 100,000 person-years at risk in a cohort of asbestos workers compared to those in other blue collar occupations (Source: Steenland and Thun 1986)

	Rate in smokers (*RR*)	Rate in non-smokers	Rate ratio
Asbestos	935.8 ($RR_{11} = 32.7$)	500.5 ($RR_{01} = 17.5$)	1.9
Non-asbestos	199.5 ($RR_{10} = 7.0$)	28.6 ($RR_{00} = 1.0$)	7.0
Rate ratio	4.7	17.5	
Rate difference	736.3	471.9	

further often reasonable assumptions of no prevention by either factor (causal monotonicity), non-additivity will also signal the presence of synergistic sufficient-cause interactions (VanderWeele 2009a).

It should be noted that these concepts of no biological interaction are distinct from certain other biological concepts of no interaction. For example, some authors give a definition in which two factors have biologically independent effects "if the qualitative nature of the mechanism of action of each is not affected by the presence or absence of the other" (Siemiatycki and Thomas 1981). Unfortunately, this concept does not lead to an unambiguous definition of interdependent effects and thus does not produce clear analytical implications. In contrast, under the sufficient-cause and potential-outcome (counterfactual) conceptualizations cited above, the given biological model is itself evaluated in terms of the co-participation of factors in a sufficient cause or in terms of modification of individual response. For example, two factors which act at different stages of a multistage process have interdependent effects because they are joint components of at least one sufficient cause. This occurs irrespective of whether they affect each other's qualitative mechanism of action. They may not even need to be in the body at the same time, for example, if one risk factor affects an early stage of the process and the other affects a much later stage. Thus, whether or not two factors are synergistic or co-participate in a sufficient cause is a completely different issue from whether they affect each other's mechanism of action and can be empirically tested under a variety of scenarios (Rothman et al. 2008, Chap. 16; VanderWeele and Robins 2007, 2008; VanderWeele 2009a, b; VanderWeele et al. 2010).

18.3.4 Additive and Multiplicative Models

The sufficient-component and potential-outcome definitions of interaction (co-participation in a sufficient cause or change in response schedule) are attractive because they are based on an explicit causal model that leads to an unambiguous definition of independence of effects and because they lead to the additive model as the baseline for assessing interactions, just as obtained through public health (cost-benefit) considerations (Rothman et al. 1980, 2008, Chap. 5). However, the analytical implications of these concepts are not straightforward, since assessing independence of effects is usually only one of the analytical goals of an epidemiological study. There are several other considerations which often favor the use of multiplicative models.

One is that multiplicative models have convenient statistical properties. Estimation in non-multiplicative models may have problems of convergence, and inference based on the asymptotic standard errors may be flawed unless the study size is very large (Moolgavkar and Venzon 1987). Another is that, if it is desired to keep interaction (effect-measure modification, corresponding to product terms in a regression model) to a minimum, then a multiplicative model is often most effective. It is not uncommon for joint effects to appear closer to multiplicative than to additive (Saracci 1987). In this situation, there may be less masking of heterogeneity in calculating an overall rate ratio than in calculating an overall rate difference. Although departures from additivity need not be multiplicative (Selikoff et al. 1980; Saracci 1987), multiplicative-model summaries tend to be closer to a population-average (standardized) measure than are additive-model summaries (Greenland and Maldonado 1994). Finally, additive-risk models are not identical to additive relative-risk models when the model includes terms for confounder adjustment; unfortunately, in typical case-control studies, only the latter models can be fit, thus rendering it difficult or impossible to make unconfounded assessments of risk additivity (Greenland 1993a, b). In contrast, departures from multiplicativity can be assessed in the same fashion from cohort and case-control data.

18.3.5 Detecting Effect-Measure Modification

Determining whether or not a factor is an effect-measure modifier is often done by estimating an effect measure (e.g., relative risk) for the exposure of interest separately for each level of the presumed effect modifier and testing for equality of these measures across the modifier strata (Rothman et al. 2008, Chap. 15). This approach lacks power, however, and so it can be quite misleading to conclude modification is absent just because the test yields a large *p*-value. Because of such power problems and other problems due to sample size limitations, when there are multiple possible effect-measure modifiers (such as age, ethnicity, gender, or previous employment in a hazardous industry), effect modification is usually examined for each potential modifying variable separately or else through use of modeling methods that allow continuous modification by quantitative variables such as age. Prior selection of potential effect modifiers of greatest interest can simplify the task. Then, assessing effect-measure modification for a subset of modifying variables might be carried out, with adjustment made for other variables.

A major obstacle to modification as well as confounding assessment is misclassification. Misclassification of any of the variables in the analysis (whether the exposure, disease, confounder, or modifier) can make a measure appear to vary across strata when in reality it does not or make it appear nearly constant when in reality it does vary (Greenland 1980). Similarly, measurement errors can spuriously create or mask the need for product terms ("interactions") in a statistical model (Greenland 1993b; cf. chapter ►Regression Methods for Epidemiological Analysis of this handbook). In an analogous fashion, variation in a measure across strata may be spuriously created or masked by differences in other biases (such as residual confounding or selection bias) across strata. Again, such problems can be explored using sensitivity analysis.

Conventional statistical analysis strategies often assume that it is not appropriate to calculate an overall effect estimate if interaction is present. However, this principle is commonly ignored if the difference in stratum-specific effect estimates is not too great. In fact standardized rate ratios have been developed for precisely this situation, and will consistently estimate meaningful epidemiological parameters even under heterogeneity (Greenland 2004; Rothman et al. 2008, Chap. 15). And again, rate ratios estimated from multiplicative models often approximate these standardized ratios (Greenland and Maldonado 1994).

As mentioned above, concluding that there is no effect-measure modification because the p-value is high can be misleading. Most studies are not designed to examine modification, and as such, may have inadequate study sizes within strata of an effect-measure modifier to permit a useful statistical interpretation. Presentation of stratum-specific effect estimates and their confidence intervals can help to give a picture of whether the data allow any inference about effect-measure modification. Formal statistical tests may be most useful in situations where prior information suggests likely forms of effect modification (e.g., a harmful effect would only be anticipated among smokers) and the study is intentionally designed to accommodate an analysis of effect modification (e.g., sufficient numbers of smokers and non-smokers are selected).

Some authors (e.g., Kleinbaum et al. 1982) have developed modeling strategies in which the first step of an analysis involves testing for effect-measure modification, where the latter is represented by product terms in the model. In the most extreme application, this involves including all possible two-factor (and even three-factor) product terms in a preliminary model and retaining in subsequent analyses all products (and related lower-order terms) that meet the inclusion criterion (which might be having a p-value below a certain cut-off, such as 0.10, or having a point estimate larger than a particular magnitude). This approach often results in complex models with numerous product terms, which may lead to problems of convergence, bias in the parameter estimates due to data sparsity, and difficulties in interpretation.

In fact, there is no logical necessity for the assessment of effect-measure modification as the first step in an analysis, and there are several reasons why it can be preferable to evaluate confounding before considering modification. One reason is that the initial aim of most analyses is to determine if there is any overall effect of exposure. It is necessary to control confounding to do this, but it is not essential to evaluate modification when doing so. Although harmful effects in one stratum and protective effects in another stratum may yield an overall null effect, this phenomenon is presumably rare. A routine search may yield a high percentage of false-positives; on the other hand, if there were a relevant a priori hypothesis, then it would be appropriate to calculate stratum-specific effect estimates irrespective of the value of the summary effect estimate.

Another reason to begin an analysis with confounding evaluation is that inclusion of extra stratification variables or extra product terms involving the main exposure complicates confounder assessment. With extra strata or terms, changes in either stratum-specific or in summary fitted measures must be examined; the stratum-specific measures may be numerous and unstable, and the summary of these measures can be difficult to construct from a fitted model that has product terms involving the exposure, see (Rothman et al. 2008, Chap. 21) and (Greenland 2004) for example formulas. Ratio measures constructed from multiplicative models that omit product terms are often a reasonable approximation to the formally correct and more complicated measures that incorporate the terms, and so can be adequate for confounding evaluation (Greenland and Maldonado 1994).

Even if subsequent analyses concentrate on specific subgroups, it may be preferable to evaluate confounding in the whole data set, since this provides the greatest precision. If a factor is a confounder overall, then it is a risk factor, and is also associated with exposure. Thus it is necessarily a confounder in some specific subgroups, and there may be little loss of precision from control in any subgroups in which it is not a confounder (although this cannot be guaranteed). Hence, it may be preferable to evaluate confounding first, and then adjust for the same confounders in each subgroup analysis.

Some qualifications should be noted. First, confounding may be evaluated purely on contextual (subject-matter) considerations and, as mentioned earlier, has an inescapable a priori (causal) component in observational studies. Because of this causal component, purely statistical selection procedures such as stepwise regression can be even more misleading for confounder selection than they are in pure prediction problems (Greenland and Neutra 1980; Rothman et al. 2008, Chap. 12). Second, the entire selection process and the attendant problems can be avoided by switching to hierarchical regression methods (Greenland 2000b, c, 2008; Rothman et al. 2008, Chap. 21), which we discuss further below. For general principles of data analysis, we refer to chapter ▶Analysis of Continuous Covariates and Dose-Effect Analysis of this handbook and Chaps. 10 and 13 of Rothman et al. (2008).

18.3.6 Assessment of Joint Effects

The above considerations imply an apparent dilemma of how to conduct an analysis that combines the advantages of ratio measures of effect with the assessment of additivity, as implied by no interaction in the potential-outcomes framework. If an excess risk is found (and assumed to be causal) then attention shifts to elaborating the nature of the effect. This naturally comes toward the end of the formal presentation of the findings. Typically, the last few tables of a manuscript might examine the joint effects of the main exposure with other factors of interest, and the discussion might relate these findings to current etiological knowledge. As noted above, it often suffices to evaluate only those joint effects for which there is an a priori reason for interest.

As an example, when studying asbestos and lung cancer, interaction with smoking might be expected given the powerful effects of smoking. To examine the latter interaction, relative risks might be presented for smoking (in non-asbestos workers), asbestos exposure (in non-smokers) and exposure to both factors, relative to persons exposed to neither factor. These relative risks would be adjusted for all other factors (e.g., age) that are potential confounders, but not of immediate interest as effect modifiers. The relevant table (e.g., Table 18.3) can be derived from any form of model, including the statistically convenient multiplicative models, by including product terms as appropriate or by including separate variables for the mutually exclusive categories of exposure (asbestos alone, smoking alone, both).

The estimation of separate and joint effects may be difficult when the factors of interest are closely correlated. However, when it is feasible, this approach combines useful features of multiplicative models and additive interaction assessment; it also permits readers with other concepts of interaction to draw their own conclusions. Consider again the data in Table 18.3 on asbestos exposure, cigarette smoking, and lung cancer. The rate-ratio estimates (adjusted for age and calendar period) are 7.0 for asbestos exposure alone, 17.5 for smoking alone, and 32.7 for the joint effect of both exposures. Thus, the joint effect of asbestos and smoking is more than additive (the joint effect is 32.7 times, whereas it would be $1 + (7.0 - 1) + (17.5 - 1) = 23.5$ if it were additive).

This result is consistent with the hypothesis that asbestos and smoking are joint components in at least one sufficient cause; it might even be argued that the observed non-additivity refutes the hypothesis that asbestos and smoking never biologically interact (assuming as usual that there is no residual confounding or bias). If the joint effect were the sum of the separate effects, the result would have favored the hypothesis that they are not joint components of a sufficient cause and do not compete for a common pool of susceptibles. However, the latter interpretation is more restricted, since additivity could arise if two factors were components of the same sufficient cause, but also had antagonistic or competitive effects that balanced their synergistic effects. Thus, even in ideal circumstances, additivity does not definitively refute the hypothesis that asbestos and smoking interact biologically in some people (Greenland and Poole 1988, Chap. 5; Rothman et al. 2008). Furthermore, without further assumptions, sufficient-cause interactions may be absent and yet there may be departures from additivity (VanderWeele 2009a, b).

Nonetheless, if it is provisionally accepted that smoking and asbestos do act together in a sufficient cause of lung cancer, then attention may shift to elaborating the effect with mathematical models deduced from biological models of the interaction. For example, Doll and Peto (1978) suggested that smoking acts at both an early stage (probably the second) and the penultimate (fifth) stage of a six-stage carcinogenic process. Asbestos appears to act at one of the later stages, probably the fourth or fifth (Pearce 1988). If asbestos acted at the same late stage as smoking, then it could be expected that its effect would add onto the late-stage effect of smoking and multiply the early-stage effect of smoking. The resulting joint effect would be intermediate between additive and multiplicative. This pattern has been observed in several studies (Selikoff et al. 1980) although there are, of course, other models which predict the same result (Saracci 1987).

When interaction evaluation occurs as the last stage of an analysis, the routine evaluation (screening) of a large number of joint effects increases the number of tables but does not necessarily complicate other aspects of the presentation (Pearce 1989). It does however raise a number of statistical problems which have been the subject of much controversy and research. The first, lesser known problem is that exposure effect estimates may be biased away from the null when too many terms (such as product terms) are entered into a risk or rate regression (Greenland et al. 2000). The second is the multiple-comparisons problem. Although many epidemiologists have denied that such problems exist (e.g., Rothman 1990), their focus concerned situations in which despite many comparisons, the investigator was interested in just one or a few exposure-disease relations. Nonetheless, screening a large number of effects (whether main effects or product terms) implies interest in many relations and raises the issues of how one deals with the instability of the estimates and the high probability that some of the estimates are large simply because of large random errors (Greenland and Robins 1991; Greenland 1993c, 2000b, c; Steenland et al. 2000). Classical multiple-comparison procedures can be quite misleading, however, because they make no attempt to account for false-negative error (in fact they inflate it tremendously) and are arguably inferior to making no adjustment at all if one is less concerned about false-negatives than false-positives.

An analytical solution to both problems is to employ hierarchical modeling methods (also known as multilevel methods, penalized estimation, random-coefficient regression, shrinkage estimation, Stein estimation, empirical-Bayes regression, and semi-Bayes regression) (Greenland and Robins 1991; Greenland 1993c, 2000b, c; Steenland et al. 2000; Rothman et al. 2008, Chap. 21). Such methods are demonstrably superior to either extreme (of no adjustment versus classical adjustment) in these situations, as shown by theory, simulations, and performance in real epidemiological examples (Efron and Morris 1977; Greenland 1993c, 2000b, c, 2008; Steenland et al. 2000; Witte et al. 2000). Furthermore, these methods can also be applied to control for multiple confounders in place of confounder selection methods (Greenland 2000c) and can be carried out with standard software (Greenland 2008; Witte et al. 2000). Nonetheless, no analytical method can overcome the fundamental limits of precision and validity that afflict typical epidemiological data; such limits will usually mean that only substantial modification and interaction will be detectable in typical epidemiological research (Greenland 1993b, 2009b).

18.4 Conclusions

Confounding occurs when the exposed and non-exposed subpopulations of the source population have different background disease risks. When we make a comparison of the frequency of disease in the exposed and non-exposed groups, we would ideally wish to be able to assume that the disease frequency in the non-exposed group provides a valid estimate of what the disease frequency would have been in the exposed group if it had not been exposed. If this assumption is incorrect, that is, if the exposed and non-exposed groups would have had different disease frequencies in the counterfactual situation in which the exposed group had not been exposed, then we say that the comparison of the exposed and non-exposed groups is confounded. A related concept is that the exposed and non-exposed group are not "exchangeable," in that the estimated effects would have been different if the exposed group had not been exposed and the non-exposed group had been exposed (i.e., if the exposure status of the subjects had been exchanged).

Interaction usually means that the exposure effect on disease risk varies by some other factor. In other words, in order to estimate the effect of exposure, we must first know whether or not another factor is present (or the level of this other factor). This idea turns out to subsume two separate concepts: *effect-measure modification* (statistical interaction) and *biological interaction*. When considering an exposure that has an effect, all other causal factors will modify either the rate ratio or the rate difference, and uniformity over one measure implies non-uniformity over the other, for example, an apparent additive joint effect implies a departure from a multiplicative model. Effect-measure modification is algebraically equivalent to the definition of "interaction" used in most statistics books and programs; it refers to variation in a population measure of effect and corresponds to the need for a product term in a model. In contrast, biological interaction refers to effects in individuals; for purely causal factors, its absence implies absence of risk-difference modification, for example, if there is no biological interaction between causal factors A and B, then the risk difference due to factor A will be independent of whether or not factor B is present or absent, and vice versa.

In the simple case of a dichotomous main exposure (e.g., asbestos exposure), a dichotomous health outcome (e.g., lung cancer), and another categorized exposure (e.g., smoking vs. non-smoking), assessment of confounding involves stratifying on the potential confounder and assessing whether the stratum-specific effect estimates are similar to the (crude) overall effect estimate. For example, we may ask "how close are the relative risks in smokers and non-smokers (or a summary of these stratum-specific effect estimates) to the relative risk estimated when smoking is ignored?" In contrast, assessment of effect-measure modification involves assessment of how the stratum-specific effect estimates compare with each other, for example, how does the relative risk in smokers compare with the relative risk in non-smokers? Effect-measure modification is thus often confused with confounding because in this simple situation they are both assessed by stratification. However, confounding and interaction are completely different concepts. A factor may be a source of confounding or effect-measure modification, or both, or neither.

Acknowledgements The authors are grateful to Katherine J. Hoggatt and the editors for numerous helpful comments that improved this chapter. Funding for Neil Pearce's salary is from a Programme Grant from the Health Research Council of New Zealand.

References

Armstrong BG (1998) Effect of measurement error on epidemiological studies of environmental and occupational exposures. Occup Environ Med 55:651–656

Armstrong B, White E, Saracci R (1992) Principles of exposure measurement in epidemiology. Oxford University Press, New York

Axelson O (1978) Aspects on confounding in occupational health epidemiology. Scand J Work Environ Health 4:85–89

Axelson O (1989) Confounding from smoking in occupational epidemiology. Br J Ind Med 46:505–507

Axelson O, Flodin U, Hardell L (1982) A comment on the reference series with regard to multiple exposure evaluations in a case-referent study. Scand J Work Environ Health 8(Suppl 1):15–19

Axelson O, Steenland K (1988) Indirect methods of assessing the effects of tobacco use in occupational studies. Am J Ind Med 13:105–118

Bross IDJ (1967) Pertinency of an extraneous variable. J Chronic Dis 20:487–495

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu C (2006) Measurement error in non-linear models. Chapman and Hall, Boca Raton

Checkoway H, Pearce N, Kriebel D (2004) Research methods in occupational epidemiology. Oxford University Press, New York

Checkoway H, Waldman GT (1985) Assessing the possible extent of confounding in occupational case-referent studies. Scand J Work Environ Health 11:131–133

Cole SR, Hernan MA (2002) Fallibility in estimating direct effects. Int J Epidemiol 31:163–165

Copeland KT, Checkoway H, McMichael AJ, Holbrook RH (1977) Bias due to misclassification in the estimation of relative risk. Am J Epidemiol 105:488–495

Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. J Natl Cancer Inst 22: 173–203

Doll R, Peto R (1978) Cigarette smoking and bronchial carcinoma: dose and time relationships among regular smokers and lifelong nonsmokers. J Epidemiol Community Health 32:303–313

Efron B, Morris C (1977) Stein's paradox in statistics. Sci Am 236:119–127

Espeland M, Hui SL (1987) A general approach to analyzing epidemiologic data that contain misclassification errors. Biometrics 43:1001–1012

Fingerhut MA, Halperin WE, Marlow DA, Piacitelli LA, Honchar PA, Sweeney MH, Greife AL, Dill PA, Steenland K, Suruda AJ (1991) Cancer mortality in workers exposed to 2,3,7,8-tetrachlorodibenzo-p-dioxin. N Engl J Med 324:212–218

Flanders WD, Khoury MJ (1990) Indirect assessment of confounding: graphic description and limits on effect for adjusting for covariates. Epidemiology 1:239–246

Fox MP, Lash TL, Greenland S (2005) A method to automate probabilistic sensitivity analyses of misclassified binary variables. Int J Epidemiol 34:1370–1377

Greenland S (1980) The effect of misclassification in the presence of covariates. Am J Epidemiol 112:564–569

Greenland S (1988) Variance estimation for epidemiologic effect estimates under misclassification. Stat Med 7:745–757

Greenland S (1993a) Additive-risk versus additive relative-risk models. Epidemiology 4:32–36

Greenland S (1993b) Basic problems in interaction assessment. Environ Health Perspect 101:59–66

Greenland S (1993c) Methods for epidemiologic analyses of multiple exposures: a review and a comparative study of maximum-likelihood, preliminary testing, and empirical-Bayes regression. Stat Med 12:717–736

Greenland S (1996) Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. Epidemiology 7:498–501

Greenland S (2000a) An introduction to instrumental variables for epidemiologists. Int J Epidemiol 29:722–729

Greenland S (2000b) Principles of multilevel modelling. Int J Epidemiol 29:158–167

Greenland S (2000c) When should epidemiologic regressions use random coefficients? Biometrics 56:915–921

Greenland S (2001) Ecologic versus individual-level sources of confounding in ecologic estimates of contextual health effects. Int J Epidemiol 30:1343–1350

Greenland S (2003) The impact of prior distributions for uncontrolled confounding and response bias: a case study of the relation of wire codes and magnetic fields to childhood leukemia. J Am Stat Assoc 98:47–54

Greenland S (2004) Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. Am J Epidemiol 160:301–305

Greenland S (2005) Multiple-bias modeling for observational studies. J R Stat Soc A 168:267–308

Greenland S (2008) Variable selection and shrinkage in the control of multiple confounders (invited commentary). Am J Epidemiol 167:523–529, Erratum: pp 1142

Greenland S (2009a) Bayesian perspectives for epidemiologic research. III. Bias analysis via missing-data methods. Int J Epidemiol 38:1662–1673

Greenland S (2009b) Interactions in epidemiology: relevance, identification, and estimation. Epidemiology 20:14–17

Greenland S, Kleinbaum D (1983) Correcting for misclassification in two-way tables and matched-pair studies. Int J Epidemiol 12:93–97

Greenland S, Maldonado G (1994) The interpretation of multiplicative model parameters as standardized parameters. Stat Med 13:989–999

Greenland S, Neutra RR (1980) Control of confounding in the assessment of medical technology. Int J Epidemiol 9:361–367

Greenland S, Pearl J, Robins JM (1999a) Causal diagrams for epidemiologic research. Epidemiology 10:37–48

Greenland S, Poole C (1988) Invariants and noninvariants in the concept of interdependent effects. Scand J Work Environ Health 14:125–129

Greenland S, Robins JM (1985) Confounding and misclassification. Am J Epidemiol 122:495–506

Greenland S, Robins JM (1986) Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 15:413–419

Greenland S, Robins JM (1991) Empirical-Bayes adjustments for multiple comparisons are sometimes useful. Epidemiology 2:244–251

Greenland S, Robins JM (2009) Identifiability, exchangeability, and confounding revisited (invited paper). Epidemiol Perspect Innov (online journal) 6:article 4

Greenland S, Robins JM, Pearl J (1999b) Confounding and collapsibility in causal inference. Stat Sci 14:29–46

Greenland S, Schwartzbaum JA, Finkle WD (2000) Problems due to small samples and sparse data in conditional logistic regression analysis. Am J Epidemiol 151:531–539

Gustafson P (2003) Measurement error and misclassification in statistics and epidemiology. Chapman and Hall, Boca Raton

Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA (2002) Causal knowledge as a prerequisite for confounding evaluation. Am J Epidemiol 155:176–184

Kleinbaum D, Kupper LL, Morgenstern H (1982) Epidemiologic research. Principles and quantitative methods. Lifetime Learning, Belmont

Kogevinas M, Pearce N, Susser M, Boffetta P (1997) Social inequalities and cancer. In: Kogevinas M, Pearce N, Susser M, Boffetta P (eds) Social inequalities and cancer. IARC, Lyon, pp 1–15

Kriebel D, Zeka A, Eisen EA, Wegman DH (2004) Quantitative evaluation of the effects of uncontrolled confounding by alcohol and tobacco in occupational cancer studies. Int J Epidemiol 33(5):1040-1045

Kupper LL, Hogan MD (1978) Interaction in epidemiologic studies. Am J Epidemiol 108:447–453

Lash TL, Fox MP, Fink AK (2009) Applying quantitative bias analysis to epidemiologic data. Springer, New York

Maldonado GM, Greenland S (1993) A simulation study of confounder-selection strategies. Am J Epidemiol 138:923–936

Maldonado GM, Greenland S (2002) Estimating causal effects (with discussion). Int J Epidemiol 31:421–438

Marshall JR, Hastrup JL (1996) Mismeasurement and the resonance of strong confounders: uncorrelated errors. Am J Epidemiol 143:1069–1078

Marshall JR, Hastrup JL (1999) Mismeasurement and the resonance of strong confounders: correlated errors. Am J Epidemiol 150:88–96

Miettinen OS (1974) Confounding and effect-modification. Am J Epidemiol 100:350–353

Miettinen OS, Cook EF (1981) Confounding: essence and detection. Am J Epidemiol 114:593–603

Moolgavkar SH, Venzon DJ (1987) General relative risk regression models for epidemiologic studies [comment]. Am J Epidemiol 126:949–961

Pearce N (1988) Multistage modelling of lung cancer mortality in asbestos textile workers. Int J Epidemiol 17:747–752

Pearce N (1989) Analytical implications of epidemiological concepts of interaction. Int J Epidemiol 18:976–980

Pearce N, Checkoway H (1988) Case-control studies using other diseases as controls: problems of excluding exposure-related diseases [comment]. Am J Epidemiol 127:851–856

Pearl J (2009) Causality: models, reasoning and inference, 2nd edn. Cambridge University Press, Cambridge

Robins J (1989) The control of confounding by intermediate variables. Stat Med 8:679–701

Robins JM (2001) Data, design, and background knowledge in etiologic inference. Epidemiology 12:550–560

Robins JM, Greenland S (1992) Identifiability and exchangeability for direct and indirect effects. Epidemiology 3:143–155

Robins JM, Greenland S (1994) Adjusting for differential rates of prophylaxis therapy for PCP in high-dose versus low-dose AZT treatment arms in an AIDS randomized trial. J Am Stat Assoc 89:737–749

Robins JM, Hernan MA, Brumback B (2000) Marginal structural models and causal inference in epidemiology. Epidemiology 11:550–560

Robins JM, Morgenstern H (1987) The foundations of confounding in epidemiology. Comput Math Appl 14:869–916

Rothman KJ (1976) Causes. Am J Epidemiol 104:587–592

Rothman KJ (1990) No adjustments are needed for multiple comparisons. Epidemiology 1:43–46

Rothman KJ, Greenland S, Lash TL (2008) Modern epidemiology. LWW, Philadelphia

Rothman KJ, Greenland S, Walker AM (1980) Concepts of interaction. Am J Epidemiol 112: 467–470

Saracci R (1987) The interactions of tobacco smoking and other agents in cancer etiology. Epidemiol Rev 9:175–193

Savitz DA, Baron EA (1989) Estimating and correcting for confounder misclassification. Am J Epidemiol 129:1062–1071

Schlesselman JJ (1978) Assessing effects of confounding variables. Am J Epidemiol 99:3–8

Schlesselman J (1982) Case-control studies: design, conduct, analysis. Oxford University Press, New York

Selikoff IJ, Seidman H, Hammond EC (1980) Mortality effects of cigarette smoking among amosite asbestos factory workers. J Natl Cancer Inst 65:507–513

Siemiatycki J, Thomas DC (1981) Biological models and statistical interactions: an example from multistage carcinogenesis. Int J Epidemiol 10:383–387

Siemiatycki J, Wacholder S, Dewar R, Wald L, Bégin D, Richardson L, Rosenman K, Gérin M (1988) Smoking and degree of occupational exposure: are internal analyses in cohort studies likely to be confounded by smoking status? Am J Ind Med 13:59–69

Sommer A, Zeger SL (1991) On estimating efficacy from clinical trials. Stat Med 10:45–52

Steenland K, Beaumont J, Halperin WE (1984) Methods of control for smoking in occupational cohort mortality studies. Scand J Work Environ Health 10:143–149

Steenland K, Bray I, Greenland S, Boffetta P (2000) Empirical Bayes adjustments for multiple results in hypothesis-generating or surveillance studies. Cancer Epidemiol Biomark Prev 9:895–903

Steenland K, Greenland S (2004) Monte-Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. Am J Epidemiol 160:384–392

Steenland K, Thun M (1986) Interaction between tobacco smoking and occupational exposures in the causation of lung cancer. J Occup Med 28:110–118

Thomas DC, Greenland S (1983) The relative efficiencies of matched and independent sample designs for case-control studies. J Chronic Dis 36:685–697

Thomas D, Stram D, Dwyer J (1993) Exposure-measurement error: influence on exposure-disease relationships and methods of correction. Annu Rev Public Health 14:69–93

VanderWeele TJ (2009a) Sufficient cause interactions and statistical interactions. Epidemiology 20:6–13

VanderWeele TJ (2009b) On the distinction between interaction and effect modification. Epidemiology 20:863–871

VanderWeele TJ, Robins JM (2007) The identification of synergism in the sufficient-component cause framework. Epidemiology 18:329–339

VanderWeele TJ, Robins JM (2008) Empirical and counterfactual conditions for sufficient-cause interactions. Biometrika 95:49–61

VanderWeele TJ, Vansteeelandt S, Robins JM (2010) Marginal structural models for sufficient cause interactions. Am J Epidemiol 171:506–514

Walter SD, Holford TR (1978) Additive, multiplicative, and other models for disease risks. Am J Epidemiol 108:341–346

Weinberg CR (1993) Toward a clearer definition of confounding. Am J Epidemiol 137:1–8

Witte JS, Greenland S, Kim LL, Arab LK (2000). Multilevel modeling in epidemiology with GLIMMIX. Epidemiology 11:684–688

Yanagawa T (1984) Case-control studies: assessing the effects of a confounding factor. Biometrika 71:191–194

Sensitivity Analysis and Bias Analysis

19

Sander Greenland

Contents

S. Greenland
Department of Epidemiology, School of Public Health, University of California, Los Angeles, CA, USA

Department of Statistics, UCLA College of Letters and Science, Los Angeles, CA, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_60,

19.1 Introduction

Over recent decades recognition has grown that the conventional statistical models used to analyze epidemiological data cannot be reasonably claimed to be correct in the way most textbooks treat them to be. In particular, conventional models for epidemiological data-generating processes cannot be credibly taken to represent targets of primary scientific interest. For example, a logistic model for the regression of an observed disease indicator on covariate measurements would only rarely correspond closely to the causal effects on disease of the risk factors represented by the measurements. The discrepancies between the statistical model parameters and the underlying target effects are often called systematic errors, biases, or bias sources. Large biases undermine the interpretation of both frequentist statistics (such as confidence intervals) and Bayesian statistics (such as posterior intervals).

The present chapter provides an overview of one useful family of approaches to analyzing and adjusting for bias that cannot be controlled by familiar stratification or regression adjustments on measured variables. It provides basic descriptions of familiar bias problems in the case of binary variables in 2 × 2 tables, with some simple illustrations. It then provides discussions of the bias problem in terms of statistical theory, compares ordinary sensitivity analysis to probabilistic bias analysis, and concludes by considering when these analyses might be helpful or essential.

An introduction to basic Bayesian concepts (e.g., Greenland 2008 and chapter ▶Bayesian Methods in Epidemiology of this handbook) is helpful in understanding concepts of bias analysis before reading the present chapter. More extensive coverage and examples of basics including computational details can be found in Fox et al. (2005), Greenland and Lash (2008), Greenland (2009a), and Lash et al. (2009), with software implementations in Steenland and Greenland (2004, appendices), Fox et al. (2005), Orsini et al. (2008), and Lash et al. (2009), while general theories can be found in Vansteelandt et al. (2006) and Greenland (2005, 2009b), among others. Throughout, variables will be denoted by capital letters, while unspecified values for a variable will be denoted by lowercase letters; for example, X may represent an exposure variable in which case x will represent a possible but unspecified value for X.

19.2 Fundamentals of Methodological Biases in 2 × 2 Table Analysis

Table 19.1 displays data on mother's report of antibiotic use during pregnancy (coded $X = 1$ yes, 0 no) and subsequent sudden infant death syndrome (SIDS, coded $Y = 1$ yes, 0 no) from a multicenter case-control study (Kraus et al. 1989); a plus sign ("+") in a subscript indicates summation over that subscript. Given the rarity of SIDS, the population risk ratio (relative risk) comparing newborns exposed to antibiotics in utero to those unexposed is approximated by the corresponding

Table 19.1 Data from case-control study of SIDS (Kraus et al. 1989). X indicates maternal recall of antibiotic use during pregnancy, and Y indicates SIDS ($Y = 1$ for cases, $Y = 0$ for controls). A plus sign ("+") in a subscript indicates summation over that subscript

	$X = 1$	$X = 0$	Totals
$Y = 1$	$A_{+11} = 173$	$A_{+01} = 602$	$A_{++1} = 775$
$Y = 0$	$A_{+10} = 134$	$A_{+00} = 663$	$A_{++0} = 797$

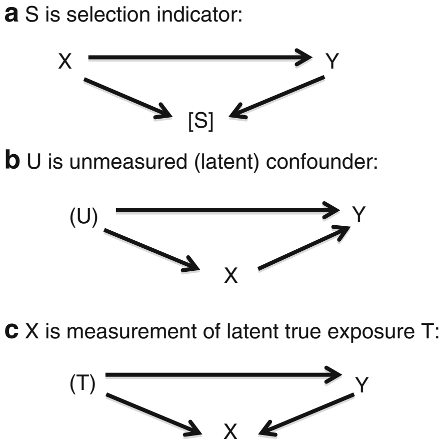

a S is selection indicator:
X
Y
[S]
b U is unmeasured (latent) confounder:
(U)
Y
X
c X is measurement of latent true exposure T:
(T)
Y
X

Fig. 19.1 Causal diagrams for three basic bias structures; *parentheses* indicate variable is unmeasured (unobserved), and *brackets* indicate observations are conditional on the variable

population odds ratio. If there were no bias, we could take this odds ratio OR_{XY} or its log, $\beta_{XY} = \ln(OR_{XY})$, as the target parameter.

The usual maximum-likelihood estimate (MLE) of $OR_{XY} = \exp(\beta_{XY})$ is the sample odds ratio $\widehat{OR}_{XY} = 173(663)/134(602) = 1.422$, with standard error for $\widehat{\beta}_{XY} = \ln(\widehat{OR}_{XY})$ of $(1/173+1/602+1/134+1/663)^{1/2} = 0.128$ and 95% confidence limits (CL) for $OR_{XY} = \exp(\beta_{XY})$ of $\exp\{\ln(1.422) \pm 1.96 \cdot 0.128\} = 1.11, 1.83$. Were there no bias, such results might be interpreted as providing an inference that OR_{XY} is above 1 but below 2. Nonetheless, experienced epidemiologists know better than to make this interpretation without cautions about common sources of bias: selection bias, uncontrolled confounding, and misclassification. As is traditional, the term "selection bias" will refer to non-random (differential or systematic) non-response and loss as well as non-random selection into the analysis.

Figure 19.1 shows the simplest representations of these three bias sources using causal directed acyclic graphs (causal diagrams), in which the arrows represent direct effects of one variable on another. With S denoting the selection indicator, the selection-bias problem (Fig. 19.1a) can be described by saying that we see the relation of X to Y conditional on $S = 1$ but we want to see that relation unconditionally (without regard to S). The confounding problem (Fig. 19.1b) can be described by saying that we see the unconditional relation of X to Y but we want to see that relation conditional on U. The misclassification problem (Fig. 19.1c) can be described by saying that we see the relation of X to Y but

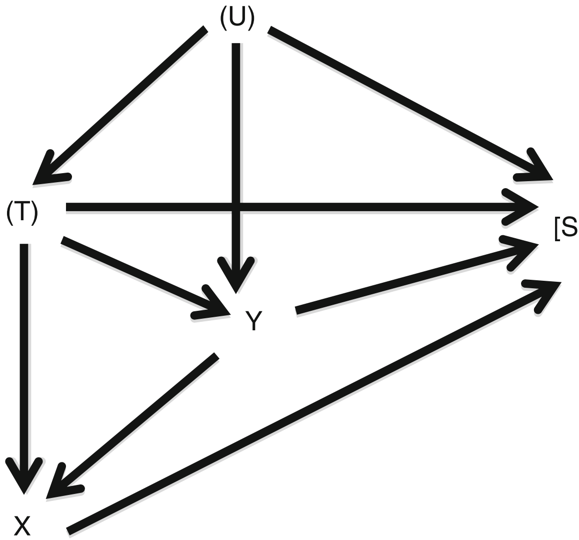


Fig. 19.2 Causal diagram combining selection, confounding, and measurement

Table 19.2 Imputed complete-population data table from SIDS study when S indicates selection into study, N_{xy} = population total with $X = x, Y = y$, and $\pi_{sxy} = \Pr(S = s|X = x, Y = y)$. Column totals form the target table

	$X = 1$		$X = 0$	
	$Y = 1$	$Y = 0$	$Y = 1$	$Y = 0$
$S = 1$	$173 = N_{11}\pi_{111}$	$134 = N_{10}\pi_{110}$	$602 = N_{01}\pi_{101}$	$663 = N_{00}\pi_{100}$
$S = 0$	$N_{11}\pi_{011}$	$N_{10}\pi_{010}$	$N_{01}\pi_{001}$	$N_{00}\pi_{000}$
Totals	$N_{11} = 173/\pi_{111}$	$N_{10} = 134/\pi_{110}$	$N_{01} = 602/\pi_{101}$	$N_{00} = 663/\pi_{100}$

we want the relation of T to Y unconditionally (without regard to X). More realistic situations involving combinations of these biases can also be represented, as in Fig. 19.2, which combines all three problems. Such diagrams are recommend as a preliminary to algebraic and statistical analyses, for they can reveal biases and assumption violations that might otherwise go undetected (Glymour and Greenland 2008; Pearl 2009; see also chapter ▶Directed Acyclic Graphs of this handbook).

Tables 19.2–19.4 display the corresponding data representations of these problems when the variables are binary, using count tables to show the target table (i.e., the unobserved structure of the target population). From these tables we can draw parallels and highlight differences among the three bias situations in Fig. 19.1:

(a) For selection bias (Fig. 19.1a and Table 19.2), we are interested in the unconditional relation of X to Y unconditional on selection S. But because we see only the relation of X to Y conditional on selection ($S = 1$), we must impute the unconditional relation of X to Y using probabilities of selection given what we do see (X and Y given $S = 1$).

(b) For confounding (Fig. 19.1b and Table 19.3), we are interested in the relation of X to Y conditional on U. But because we do not see U, we must impute its values using probabilities (bets) about the values of U given what we do see (again, X and Y).

(c) For misclassification (Fig. 19.1c and Table 19.4), we are interested in the unconditional relation of the true exposure T to the outcome Y so that if we saw T, we could discard X and just look at the TY association. But because we do not see T, we must impute its values using probabilities (bets) about the values of T given what we do see (again, X and Y).

In all three cases, we need a set of four probabilities to complete the three-way table of the three variables in the graph. These probabilities are denoted by π_{1xy} in Tables 19.2–19.4, where they correspond to probabilities of $S = 1$, $U = 1$, or $T = 1$ given the observed X and Y. We need only these four probabilities because the corresponding probabilities of $S = 0$, $U = 0$, or $T = 0$ are $\pi_{0xy} = 1 - \pi_{1xy}$. Using these probabilities, Table 19.2 displays the complete distribution of those unselected as well as those selected, while Tables 19.3 and 19.4 display the complete distribution of the observed subjects under confounding and misclassification.

Were these complete distributions known, we could compute our target parameter directly from them. But they are not known with certainty, and the best we can do

Table 19.3 Imputed complete-data table from SIDS study when U is an unmeasured confounder and $\pi_{uxy} = \Pr(U = u|X = x, Y = y)$. This table is the target table

	$X = 1$		$X = 0$	
	$Y = 1$	$Y = 0$	$Y = 1$	$Y = 0$
$U = 1$	$A_{111} = 173\pi_{111}$	$A_{110} = 134\pi_{110}$	$A_{101} = 602\pi_{101}$	$A_{100} = 663\pi_{100}$
$U = 0$	$A_{011} = 173\pi_{011}$ $= 173 - A_{111}$	$A_{010} = 134\pi_{010}$ $= 134 - A_{110}$	$A_{001} = 602\pi_{001}$ $= 602 - A_{101}$	$A_{000} = 663\pi_{000}$ $= 663 - A_{100}$
Totals	173	134	602	663

Table 19.4 Imputed complete-data and target tables from SIDS study when T indicates actual antibiotic use during pregnancy and $\pi_{txy} = \Pr(T = t|X = x, Y = y)$

(a) Complete data

	$X = 1$		$X = 0$	
	$Y = 1$	$Y = 0$	$Y = 1$	$Y = 0$
$T = 1$	$A_{111} = 173\pi_{111}$	$A_{110} = 134\pi_{110}$	$A_{101} = 602\pi_{101}$	$A_{100} = 663\pi_{100}$
$T = 0$	$A_{011} = 173\pi_{011}$ $= 173 - A_{111}$	$A_{010} = 134\pi_{010}$ $= 134 - A_{110}$	$A_{001} = 602\pi_{001}$ $= 602 - A_{101}$	$A_{000} = 663\pi_{000}$ $= 663 - A_{100}$
Totals	173	134	602	663

(b) Target table (complete data collapsed over X):

	$Y = 1$	$Y = 0$
$T = 1$	$A_{1+1} = 173\pi_{111} + 602\pi_{101}$	$A_{1+0} = 134\pi_{110} + 663\pi_{100}$
$T = 0$	$A_{0+1} = 173\pi_{011} + 602\pi_{001}$ $= 775 - A_{1+1}$	$A_{0+0} = 134\pi_{010} + 663\pi_{000}$ $= 797 - A_{1+0}$
Totals	$A_{++1} = 173 + 602 = 775$	$A_{++0} = 134 + 663 = 797$

is try to adjust our inferences using any relevant background information, such as information about the π_{1xy}, their combinations, or their components. This process is sometimes called *external* or *indirect* adjustment and is a core component of bias analysis.

19.2.1 Formulas for Selection Bias

Simplifications may arise depending on the available information, assumptions, and the target parameter. For example, in the selection-bias problem, the usual case-control target would be the population XY odds ratio OR_{XY} (the odds ratio unconditional on S), but the observed sample odds ratio, $\widehat{OR}_{XY|S=\mathit{1}}$, only estimates the odds ratio among those selected, $OR_{XY|S=1}$. The selection bias in $\widehat{OR}_{XY|S=\mathit{1}}$ relative to the population OR_{XY} is thus $\text{Bias}_{S=\mathit{1}} = \pi_{111}\pi_{100}/\pi_{110}\pi_{101} = OR_{XY|S=\mathit{1}}\big/OR_{XY}$. This bias equation suggests the following approximate estimating equation to adjust $\widehat{OR}_{XY|S=\mathit{1}}$ for selection bias:

$$\begin{aligned} OR_{XY} = N_{11}N_{00}/N_{10}N_{01} &\approx [(173/\pi_{111})(663/\pi_{100})]/[(134/\pi_{110})(602/\pi_{101})] \\ &= (173 \cdot 663)/(134 \cdot 602)/(\pi_{111}\pi_{100}/\pi_{110}\pi_{101}) \\ &= \widehat{OR}_{XY|S=\mathit{1}}/\text{Bias}_{S=1}. \end{aligned}$$

The approximation of chief statistical concern arises from substituting sample numbers for population numbers.

The equation shows that we need information on only one number, $\text{Bias}_{S=1}$, rather than all four selection probabilities π_{1xy}. $\text{Bias}_{S=1}$ might be estimated as the ratio of estimates without and with information from initial non-responders on whom data was later obtained ("call-back survey" information) from studies reporting such information (e.g., Hatch et al. 2000). Even with such information, however, concerns about generalization across studies must be addressed by considering how differences between the study populations may have affected response and selection.

19.2.2 Formulas for Confounding

Suppose there is no selection bias (so that $OR_{XY|S=1} = OR_{XY}$). Consider first the odds ratio $OR_{XY|U=1}$ in the $U = 1$ stratum (which is free of confounding by U). The bias in $\widehat{OR}_{XY}$ relative to $OR_{XY|U=1}$ is $\text{Bias}_{U=1} = \pi_{110}\pi_{101}/\pi_{111}\pi_{100} = OR_{XY}/OR_{XY|U=1}$. Table 19.3 then suggests the following approximate estimating equation to adjust $\widehat{OR}_{XY}$ for confounding by U:

$$\begin{aligned} OR_{XY|U=1} &\approx [A_{111}A_{100}]/[A_{110}A_{101}] \\ &= [(173\pi_{111})(663\pi_{100})]/[(134\pi_{110})(602\pi_{101})] \\ &= (173 \cdot 663)/(134 \cdot 602)/(\pi_{110}\pi_{101}/\pi_{111}\pi_{100}) \\ &= \widehat{OR}_{XY}/\text{Bias}_{U=1} \end{aligned}$$

A parallel formula can be written for $OR_{XY|U=0}$. To simplify analysis, however, it is usually assumed that the odds ratio is a constant $OR_{XY|U}$ across U ($OR_{XY|U=0} = OR_{XY|U=1} = OR_{XY|U}$, called the odds-ratio *homogeneity* assumption). In that case, $\text{Bias}_{U=1}$ is also the bias in $\widehat{OR}_{XY}$ relative to $OR_{XY|U=0}$, $\text{Bias}_{U=0} = \pi_{010}\pi_{001}/\pi_{011}\pi_{000} = OR_{XY}/OR_{XY|U=0}$, and both can be written as $\text{Bias}_U = OR_{XY}/OR_{XY|U} \approx \widehat{OR}_{XY}/OR_{XY|U}$. Thus, to adjust $\widehat{OR}_{XY}$ toward the constant U-specific odds ratio $OR_{XY|U}$, we need only information on one number, Bias_U, rather than all four confounder-prevalence probabilities π_{1xy}.

Suppose now that the target parameter what the effect of exposure would be on the entire population (often called the total-population or *marginal* effect of exposure). Evidence suggests that estimation of $OR_{XY|U}$ is not sensitive to heterogeneity (violations of homogeneity) in typical epidemiological settings in which the disease is uncommon and neither the XY effect nor the heterogeneity is large (Greenland and Maldonado 1994). Nonetheless, the assumption should be considered critically in any application, especially if the target parameter is the effect in a subpopulation (such as the exposed). Given the assumption, we still must estimate Bias_U. The relation $\text{Bias}_U = OR_{XY}/OR_{XY|U}$ suggests using the ratio of estimates without and with U adjustment from studies reporting such estimates; again however concerns about generalization across studies must be addressed.

The above formulation was derived for only one binary confounder U and an unadjusted $\widehat{OR}_{XY}$. Nonetheless, it extends directly to multiple confounders of any form and to situations involving a partially adjusted initial estimate (e.g., an initial estimate adjusted for age and sex but not smoking habits). This can be done by allowing $\widehat{OR}_{XY}$ to represent the partially adjusted estimate and Bias_U to represent the ratio $OR_{XY|partial}/OR_{XY|full}$ between a partially adjusted measure $OR_{XY|partial}$ and a fully adjusted measure $OR_{XY|full}$; this ratio can be estimated from reports or data providing estimates with both degrees of adjustment.

19.2.3 Sensitivity Analysis and Probabilistic Bias Analysis

The process of seeing how results change under different choices for a bias parameter (such as Bias_S or Bias_U) is often called a *sensitivity analysis*. When the focus is on the degree of bias implied by each parameter choice, the process is sometimes called *bias analysis*. The term "sensitivity analysis" is also used more broadly to include examination of the impact of any change in the statistical model, such as varying the model for the regression or random error; such model-sensitivity analysis is a large topic in econometrics and engineering (Leamer 1978, 1985; Saltelli et al. 2000) but is not covered here.

For selection bias or confounding, a sensitivity analysis might comprise nothing more than noting that, based on other data, the bias factor seems likely to fall in range, e.g., 0.80–1.33, and then displaying the adjusted point estimates for bias factors in this range. But it is not clear how to gauge or combine the uncertainty this range represents with the conventional statistics. One naïve approach would expand the confidence limits of 1.11, 1.83 in the SIDS example using the extremes of the bias range, to produce limits of 1.11/1.33 = 0.83 and 1.83/0.80 = 2.29. As shown below, this approach tends to overstate the uncertainty because it combines only the extremes of bias and random error, ignoring that in the majority of possible combinations at least one of bias and random error is small or that in half the cases the two would be in opposite directions. For example, random error is just as likely to partially counterbalance bias as add to it.

These preceding observations suggest that more modest combinations of bias and random error are more probable than extremes, an intuition that is captured mathematically in *probabilistic bias analysis*, and which is also treated under related topics such as uncertainty assessment and risk analysis (Eddy et al. 1992; Lash and Fink 2003; Phillips 2003; Steenland and Greenland 2004; Greenland 2001, 2003, 2005; Greenland and Lash 2008; Lash et al. 2009; Turner et al. 2009). As a simple example, one might specify a *prior probability* that the bias factor Bias_U for confounding by U falls in a stated range like 0.80, 1.33, typically 95%. The term "prior" signals that the assigned probability may be nothing more than an educated bet based on available relevant literature, although it could be based on detailed analysis of other studies. Regardless, one can use standard lognormal formulas to combine the bias uncertainty seen in this interval and the uncertainty from random error represented by the standard error or confidence interval.

Specifically, suppose the range 0.80, 1.33 were taken to represent the 2.5th and 97.5th percentiles of a lognormal prior-probability distribution ("prior") for Bias_U. Uncertainty about $\ln(\mathrm{Bias}_U)$ would then follow a normal distribution, with mean m_U the average of the log limits: $\mathrm{m}_U = [\ln(1.33) + \ln(0.80)]/2 = 0.0310$. We can subtract this mean bias m_U from the log odds-ratio estimate $\hat{\beta}_{XY} = \ln(\widehat{OR}_{XY})$ to get a bias-adjusted estimate of

$$\hat{\beta}_{XY} - m_U = \ln(1.422) - 0.0310 = 0.320.$$

The prior standard deviation would be the width of the log-bias interval $\ln(1.33) - \ln(0.80) = 0.508$ divided by the number of standard deviations in its width, which is $2 \cdot 1.96 = 3.92$ for a normal 95% interval. We then square this ratio to get the prior log-bias variance: $v_U = (0.508/3.92)^2 = 0.0168$. This bias variance is then added to the estimated random variance $v_R = 0.128^2 = 0.0164$ of $\ln(\widehat{OR}_{XY})$ to get a total variance $v_+ = v_R + v_U = 0.0164 + 0.0168 = 0.0332$. Finally, we compute an approximate bias-adjusted 95% *posterior probability* interval

$$\exp\{0.320 \mp 1.96(0.0332)^{1/2}\} = [0.96, 1.97].$$

This interval is more narrow than in the extreme-adjusted interval of [0.83, 2.29] but is still noticeably wider than the original confidence interval [1.11, 1.83] (as well as being shifted downward by a small amount).

To adjust for selection bias in addition to confounding, we could repeat the above process by specifying a selection-bias interval, converting that to a prior mean m_S

and variance v_S for $\ln(\text{Bias}_{S=1})$, then computing the posterior interval from the adjusted estimate $\hat{\beta}_{XY} - m_S - m_U$ and the total variance $v_R + v_S + v_U$. This extension assumes that the confounding and selection bias are independent, which would not be the case in the situation depicted in Fig. 19.2, where everything including U directly affects selection S.

The above log-linear approach to selection bias and confounding appears simple in concept but raises the question of how one can justify assigning a particular prior distribution to the bias factors. One way is to base the interval estimate for $\text{Bias}_U = OR_{XY}/OR_{XY|U}$ or $\text{Bias}_{S=1} = OR_{XY}/OR_{XY|S=1}$ seen upon adjustment in a past study, then use that interval for the current adjustment (Greenland and Mickey 1988; Greenland 2003). More often however such direct bias information is not readily available or is not considered generalizable from its source.

19.2.4 Confounding Parameters

The confounding bias factor Bias_U can be expressed as a function of the confounder distribution along with measures of the association of the confounder with exposure and disease (Kitagawa 1955; Cornfield et al. 1959; Bross 1966, 1967; Schlesselman 1978; Breslow and Day 1980, Sect. 3.4; Yanagawa 1984; Gail et al. 1988; Flanders and Khoury 1990; Schneeweiss 2006; VanderWeele and Arah 2011). For example, under homogeneity the odds-ratio bias Bias_U can be written

$$\text{Bias}_U = (1 + \theta \cdot OR_{UX|Y} \cdot OR_{UY|X})(1 + \theta)/(1 + \theta \cdot OR_{UX|Y})(1 + \theta \cdot OR_{UY|X})$$

where $\theta = \pi_{100}/\pi_{000} = \pi_{100}/(1 - \pi_{100})$ is the odds of $U = 1$ when $X = Y = 0$, $OR_{UX|Y}$ is the constant UX odds ratio given Y, and $OR_{UY|X}$ is the constant UY odds ratio given X (Yanagawa 1984). It can happen that there is some generalizable background information to place priors on these three bias parameters even though there is no direct information about Bias_U. For example, if the exposure ($X = 1$) and disease ($Y = 1$) are uncommon in the general population, then the prevalence π_{100} can be estimated from population survey data on U. If the exposure is rare or its effect on Y is "weak" ($0.5 < OR_{XY|U} < 2$), then we may expect $OR_{UY|X}$ to be similar to OR_{UY}, and OR_{UY} may be available from earlier studies.

The preceding approach assumes that U is a known confounder (e.g., a smoking indicator) that was unmeasured in the study in question but has been previously identified and subject to study in relation to disease if not exposure. If instead U represents an unspecified, unknown confounder, then the entire sensitivity exercise will remain far more speculative. Nonetheless, decomposition of the bias factor can still be successful in demonstrating that only implausibly strong confounder or selection effects can account for a strong observed association. Cornfield et al. (1959) is considered a landmark study in which such an approach was used to examine claims that the smoking-lung cancer relation might be attributable to confounding.

One approach to expanding intervals to account for uncertainties about multiple bias parameters is *Monte Carlo sensitivity analysis* (MCSA, also known as Monte Carlo risk analysis) which is based on taking random draws for each component parameter in the bias adjustment. There are many ways to proceed. One simple approach assigns prior distributions to $\text{Bias}_{S=1}$, θ, $OR_{UX|Y}$, and $OR_{UY|X}$ (which may be dependent) and a normal sampling distribution to the unadjusted log odds ratio $\hat{\beta}_{XY}$ with mean equal to the point estimate ln(1.422) and standard deviation equal to $v_R^{1/2} = 0.128$. We then cyclically generate thousands of draws from each of these distributions. Let $\text{Bias}^*_{S=1}$, θ^*, $OR^*_{UX|Y}$, $OR^*_{UY|X}$, $\hat{\beta}^*_{XY}$ be the values drawn in one such cycle. At each draw we compute a Bias^*_U from θ^*, $OR^*_{UX|Y}$, and $OR^*_{UY|X}$ and then compute $\hat{\beta}^*_{adj} = \hat{\beta}^*_{XY} - \ln(\text{Bias}^*_{S=1}) - \ln(\text{Bias}^*_U)$. We may then create a histogram of the adjusted estimates from this simulation and take the 2.5th and 97.5th percentiles of the distribution of $\hat{\beta}^*_{adj}$ so generated as a simulated 95% posterior interval for the *XY* log odds ratio β_{XY}.

An alternative "bootstrap" approach to incorporating random error is to resample (with replacement) the entire data set at each cycle and recompute $\hat{\beta}^*_{XY}$ from the resampled data. This can work well with large cell counts but may slow computation considerably and may require complex corrections in small or sparse data sets in order to account for zero cells and other problems (Efron and Tibshirani 1994).

The posterior-interval interpretations given above are approximate and assume implicitly that there is no prior information on any parameter other than the bias parameters $\text{Bias}_{S=1}$, θ, $OR_{UX|Y}$, and $OR_{UY|X}$. This assumption is not realistic in practice because (apart from genes) few studies are the first to examine a given association. As a consequence, some authors call the resulting interval a "simulation interval" without specifying a statistical meaning for the interval (e.g., Lash et al. 2009). Nonetheless, a much more severe criticism applies to the conventional 95% confidence interval: It is routinely interpreted as a posterior interval for the target parameter $OR_{XY|U}$, yet this interpretation is based on the same assumption of no prior information about this target, along with the extremely unrealistic assumption that there is no bias at all (which would be represented by priors for $\text{Bias}_{S=1}$ and Bias_U that are point masses at 1).

If one wishes to employ informative priors for parameters other than the bias parameters, it will be necessary to switch to a penalized-likelihood or explicitly Bayesian analysis format in order to correctly interpret the resulting intervals as posterior probability intervals (Joseph et al. 1995; Graham 2000; Gustafson et al. 2001, 2010; Scharfstein et al. 2003; Steenland and Greenland 2004; Gustafson 2003, 2005; Chu et al. 2006; McCandless et al. 2007, 2012; Greenland 2005, 2009a, b; Molitor et al. 2009; Turner et al. 2009; Welton et al. 2009). The traditional alternative, ordinary sensitivity analysis, avoids this complication by simply displaying results for various choices of bias parameters (Copas and Li 1997; Rosenbaum 2002; Brumback et al. 2004), but can become unwieldy if the number of bias parameters is more than three (as will happen when misclassification is addressed).

For those seeking to avoid explicit prior distributions, the dimensionality problem may be addressed by estimating regions of compatibility between the data and the parameters of interest given the random-error model and sharp bounds for the bias parameters (Vansteelandt et al. 2006). These regions generalize confidence intervals to account for the bias parameters. Nonetheless, sharp bounds for the bias parameters are needed to construct the regions, and it is not clear why they provide a less arbitrary solution to the problem than do prior distributions, since the results are just as sensitive to choice of bounds as to choice of priors (Greenland 2009b).

19.3 Bias Formulas for Misclassification

Misclassification adjustment is usually much more difficult than others, in part because the target parameter is the association (such as the odds ratio OR_{TY}) between the completely unobserved true exposure indicator T and Y, not X and Y. Its complexity is sometimes overlooked, in part because it appears that the problem can be solved by simply dividing $\widehat{OR}_{XY}$ by the bias factor $\text{Bias}_M = OR_{XY}/OR_{TY}$ to adjust for the misclassification.

Unfortunately, unlike with selection bias and confounding, the bias-factor approach is not widely useful for misclassification adjustment. First, although the bias factor could be estimated as the ratio of estimates of OR_{XY} and OR_{TY} from studies reporting both estimates, such reports are very uncommon. Second, even when an estimate of Bias_M is available, there are exceptional concerns about generalization across studies: Bias_M can be highly sensitive to the prevalence of true exposure ($T = 1$), this prevalence may vary greatly across populations, and the extent of variation may be highly uncertain because exposure prevalence may have never been reported without misclassification.

19.3.1 Formulas Based on Sensitivity and Specificity

The aforementioned problems have led the misclassification literature to focus on quantities that are more often published and more transportable (generalizable) than bias factors. Let $\phi_{xty} = \Pr(X = x|T = t, Y = y)$. The *sensitivities* of X for T are then the ϕ_{11y} and the *specificities* are ϕ_{00y}, which are sometimes studied in themselves to inform later uses of X as a measurement of T. For example, patient recall of medical events or treatments (X) is often compared to medical record information (T). The latter information is of course imperfect and is quite subject to error regarding, for example, use of drugs but may be accurate enough for other exposures (e.g., implanted devices or drug *prescriptions*) to be cautiously treated as the truth or "gold standard" for those exposures. The non-differentiality assumption (independence of X and Y given T) then corresponds to $\phi_{xty} = \Pr(X = x|T = t, Y = y) = \Pr(X = x|T = t)$ so that sensitivities and specificities remain constant across Y ($\phi_{111} = \phi_{110}$ and $\phi_{001} = \phi_{000}$).

Sensitivities and specificities are often called “true positive rates” and “true negative rates,” respectively. In parallel, the complements of sensitivities and specificities,

$$\Pr(X = 0|T = 1, Y = y) = \phi_{01y} = 1 - \phi_{11y} \text{ and}$$
$$\Pr(X = 1|T = 0, Y = y) = \phi_{10y} = 1 - \phi_{00y}$$

are often called the “false-negative rates” and “false-positive rates,” respectively. Note that X being a perfect measurement of T ($X = T$) corresponds to perfect sensitivity and specificity ($\phi_{11y} = \phi_{00y} = 1$) as well as to perfect T prediction ($\pi_{11y} = \pi_{00y} = 1$).

Given a set of values for the ϕ_{xty}, we can estimate the four desired TY counts A_{t+y} from the observed XY counts A_{+xy} by solving a system of four equations (one for each X, Y combination) giving each observed XY count A_{+xy} as the sum of those with $T = 1$ and those with $T = 0$ who contribute to the count:

$$\begin{aligned} \text{observed count} = A_{+xy} &\approx \phi_{x1y}A_{1+y} + \phi_{x0y}A_{0+y} \\ &= \phi_{x1y}A_{1+y} + \phi_{x0y}(A_{++y} - A_{1+y}) \\ &= (\phi_{x1y} - \phi_{x0y})A_{1+y} + \phi_{x0y}A_{++y} \end{aligned}$$

provided all the solutions are positive. The solutions are $A_{1+y} = (A_{+xy} - \phi_{x0y}A_{++y})/(\phi_{x1y} - \phi_{x0y})$, which lead to an adjusted TY odds ratio collapsed (summed) over X:

$$\begin{aligned} \widehat{OR}_{TY} = &[(A_{+11} - \phi_{101}A_{++1})(A_{+00} - \phi_{010}A_{++0})]/ \\ &[(A_{+01} - \phi_{011}A_{++1})\ (A_{+10} - \phi_{100}A_{++0})]. \end{aligned}$$

This formula does not simplify to a simple direct adjustment to $\widehat{OR}_{XY}$ and shows that without further assumptions, we need to know four parameters to adjust for the misclassification.

Background information on the four classification parameters may sometimes be found in *validation studies*, which examine the accuracy of X as a measure of T. If the validation study is a substudy of the study under analysis (internal validation), one can and should analyze its TXY data together with the main XY data using special methods. Those methods may be viewed as missing-data techniques because with a validation substudy T becomes a partially missing variable (Little and Rubin 2002; Carroll et al. 2006), and thus may be handled by methods as simple as multiple imputation (Cole et al. 2006). A key assumption for these techniques however is that T is missing at random (MAR), which is satisfied if subjects with T measurement are a random sample of subjects within each level of the observed variables.

If the validation data are from outside the study under analysis (external validation), uncertainty about generalizability must be considered, although the validation data can help one specify a prior distribution for the classification parameters (which are bias parameters). Again, it appears that estimates of sensitivities and specificities (ϕ_{11y} and ϕ_{00y}) are more safely generalized than estimates of T-predictive probabilities (π_{1xy}) because the latter depend directly on the T distribution, which is unobserved in the analysis data and thus has unknown discrepancy from the T distribution in the validation data. An alternative approach is to use background information on sensitivities and specificities to help specify components of the π_{1xy}. This approach will be illustrated below.

19.3.2 Probabilistic Analysis Based on Sensitivity and Specificity Priors

One may parallel the earlier simulation analysis by setting priors on the sensitivities and specificities and sampling from these priors and the random errors to generate a simulation histogram of $\widehat{OR}_{TY}$ (Lash and Fink 2003; Lash et al. 2009; Fox et al. 2005; Greenland and Lash 2008). Unfortunately there are several complications that arise due to the more complex structure of misclassification adjustments. First, some combinations of ϕ_{1ty} allowed by the priors may lead to negative (and thus impossible) values for the imputed TXY counts. Such combinations are said to be *inadmissible*. If the priors chosen for the ϕ_{1ty} allow a non-negligible probability for inadmissible combinations, the resulting simulation intervals will no longer agree well with the correct Bayesian posterior intervals, whether or not the inadmissible values are discarded (Greenland 2005; MacLehose and Gustafson 2012).

Second, independent priors for the ϕ_{1ty} are grossly unrealistic. A large source of negative correlation between sensitivity and specificity is the stringency of criteria used to declare $X = 1$ (as an indicator of $T = 1$), with more stringency leading to lower sensitivity and higher specificity, and less stringency leading to higher sensitivity and lower specificity. A source of positive correlation is that use of more accurate measurement methods to create X (e.g., direct measurements) can increase both sensitivity and specificity compared to less accurate methods (e.g., self-reports). These sources of correlation need not balance to zero.

Of even greater concern, case and non-case sensitivity and specificity will be very highly positively correlated. If as usual the same method is used for measurement of cases and non-cases, the proper "null" reference point is not independence, but is instead its positive opposite, non-differentiality ($\phi_{\mathrm{xt1}} = \phi_{xt0}$). Non-differentiality implies complete dependence (and thus 100% correlation) between case and non-case sensitivity and specificity (although complete dependence is not sufficient to produce non-differentiality). For the same reasons, realistic priors for the T-predictive probabilities π_{txy} will also be highly correlated.

It is possible to construct dependent priors to account for such correlation sources, although specifying the correlations can be difficult. See Fox et al. (2005) and Greenland and Lash (2008) for details and examples.

19.3.3 Probabilistic Analysis Based on Predictive-Value Components

The predictive approach is familiar in validation-study analysis (Lyles 2002) but is also a useful alternative to using direct priors for sensitivity and specificity, because it can use instead parameters that have no admissibility restrictions and for which prior independence may not be such a severely distortive assumption (Greenland 2009a). To this end, we recast the T-predictive probabilities π_{1xy} in terms of a logistic model:

$$\pi_{1xy} = \Pr(T = 1|X = x, Y = y) = \text{expit}(\beta_T + \beta_{TX}x + \beta_{TY}y + \beta_{TXY}xy)$$

where $\text{expit}(z) = e^z/(1 + e^z)$, which can be rewritten as the odds model:

$$\begin{aligned}\pi_{1xy}/\pi_{0xy} &= \Pr(T = 1|X = x, Y = y)/\Pr(T = 0|X = x, Y = y)\\ &= \exp(\beta_T + \beta_{TX}x + \beta_{TY}y + \beta_{\mathrm{TXY}}\mathrm{xy}).\end{aligned}$$

The odds ratio relating T and X when $Y = y$ is then seen to be

$$\begin{aligned}OR_{TX|Y=y} &= (\pi_{11y}\pi_{00y})/(\pi_{10y}\pi_{01y}) = (\pi_{11y}/\pi_{01y})/(\pi_{10y}/\pi_{00y})\\ &= \exp(\beta_{TX} + \beta_{TXY}y).\end{aligned}$$

Some algebra shows that (by the usual symmetry property of odds ratios), $OR_{TX|Y=y}$ also equals $\phi_{11y}\phi_{00y}/\phi_{10y}\phi_{01y}$, which is the product of sensitivity and specificity divided by the product of false-positive and false-negative rates. This ratio is the *receiver-operating characteristic* (ROC) odds ratio; it measures the quality of X as an indicator of T, being infinite when X is perfect and 1 when X is completely worthless.

Assuming a rare disease (so that the odds ratio among non-cases $OR_{TX|Y=0}$ approximates the odds ratio OR_{TX}), a prior for $\beta_{TX} = \ln(OR_{TX|Y=0})$ can be based on expectations for the population sensitivity and specificity via the relation $OR_{TX|Y=0} = \phi_{110}\phi_{000}/\phi_{100}\phi_{010}$. For X to be a high-quality indicator of T, this ROC odds ratio must be very large (Pepe et al. 2004). For example with sensitivity 0.6 and specificity 0.8 among non-cases, $OR_{TX0} = \exp(\beta_{TX}) = (0.6/0.4)/(0.2/0.8) = 6$; 0.6 and 0.9 yield (0.6/0.4)/(0.1/0.9) = 13.5; 0.8 and 0.8 yield (0.8/0.2)/(0.2/0.8) = 16; and 0.8 and 0.9 yield (0.8/0.2)/(0.1/0.9) = 36. Thus, even for a mediocre measurement (with sensitivity of only 0.6 and specificity of only 0.8) the *TX* odds ratio among non-cases, $OR_{TX0} = \exp(\beta_{TX})$ is 6 indicating that a prior for β_{TX} should be distributed well above zero. The preceding numeric examples would be encompassed by a lognormal prior for OR_{TX0} with median 13.5 (in the center of the examples) and 95% limits of 5 and 36, which translates to a normal prior with mean ln(13.5) and variance 0.25 for β_{TX}

Table 19.5 A set of independent normal prior distributions for the coefficients in the logistic regression of T on X and Y in the antibiotic-SIDS example

	Mean	Variance	95% prior limits for
β_T	logit(0.1)	0.16	$\pi_{100} = \text{expit}(\beta_T)$: 0.05, 0.20
β_{TX}	ln(13.5)	0.25	$\exp(\beta_{TX})$: 5, 36
β_{TY}	0	0.50	$\exp(\beta_{TY})$: 0.25, 4
β_{TXY}	0	0.125	$\exp(\beta_{TXY})$: 0.5, 2

The ratio of ROC odds ratios is $\text{ROR}_{TX} = OR_{TX|Y=1}/OR_{TX|Y=0} = \exp(\beta_{TXY})$, which is a summary measure of how much better (if over 1) or worse (if under 1) X is as a T indicator among cases compared to non-cases. Under non-differential misclassification, β_{TXY} is 0 and $\exp(\beta_{TY})$ equals the target parameter OR_{TY} (although $\beta_{TXY} = 0$ is not sufficient to produce non-differentiality or $\exp(\beta_{TY}) = OR_{TY}$). Thus, unless severe non-differentiality is expected, it is reasonable to use a prior for β_{TXY} that is concentrated near zero, for example, a normal prior with mean 0 and variance 0.125, which produces 95% prior limits for ROR_{TX} of 0.5 and 2. Likewise, it would be reasonable to use a prior for β_{TY} that is similar to but somewhat wider than what would be considered a reasonable prior for the target log odds ratio $\ln(\text{OR}_{TY})$, for example, a normal prior with mean zero and variance 0.5, which produces 95% prior limits for $\exp(\beta_{TY})$ of 0.25 and 4.

In settings where recall bias is expected (differences in case and non-case recall), it is important to not be deceived into thinking such bias alone implies that ROR_{TX} is above or below 1 (β_{TXY} above or below 0). In the example in Table 19.1, the traumatic event of SIDS may trigger both true and false recall, increasing sensitivity (which increases ROR_{TX}) as well as decreasing specificity (which decreases ROR_{TX}). The net result is that uncertainty about β_{TXY} should be more symmetrically distributed around 0 than naïve reasoning might suggest.

Finally, note that $\text{expit}(\beta_T) = \pi_{100} = \Pr(T = 1|X = 0, Y = 0)$ is the probability that a "test negative" ($X = 0$) in a non-case is erroneous. For an exposure with an expected prevalence well below 50% (as was antibiotic use in unselected pregnancies during the study period) and a reasonably sensitive measure X of T, we should concentrate our prior distribution for π_{100} well below 0.5, which forces the prior for $\beta_T = \text{logit}(\pi_{100})$ to be well below 0. An example would be a normal prior with mean for β_T with mean $\text{logit}(0.1) = -2.20$ and variance 0.16, which produces 95% prior limits for π_{100} of 0.05 and 0.20.

Table 19.5 summarizes the independent normal priors suggested above for the logistic predictive-value coefficients β_T, β_{TX}, β_{TY}, β_{TXY}, which are also consistent with the limited background information about the antibiotic-SIDS relation at the time (Kraus et al. 1989). These can be used in a Monte Carlo sensitivity analysis in which at each cycle:

1. All counts are resampled from Table 19.1.
2. β^*_T, β^*_{TX}, β^*_{TY}, and β^*_{TXY} are drawn from their priors and used to compute the π^*_{txy}.
3. These π^*_{txy} are multiplied against the resampled counts to get a simulated TXY table, as in Table 19.4a.

4. The TXY table is collapsed over X, and $\widehat{OR}^*_{TY}$ is computed from the resulting TY table (Table 19.4b).

After 250,000 repetitions of sampling cycle 1–4, the 2.5th and 97.5th percentiles of the resulting $\widehat{OR}^*_{TY}$ distribution are 0.37 and 3.4, with a median of 1.2. Various Bayesian calculations with non-informative priors on all parameters besides β_T, β_{TX}, β_{TY}, and β_{TXY} yield similar results (Greenland 2009a, b). These results show that the precision of the conventional 95% confidence limits of 1.1, 1.8, if interpreted as an interval for OR_{TY}, is due to its assumption that X is a perfect measurement of $T(X = T$, i.e., $\pi_{11y} = \pi_{00y} = 1$ or $\phi_{11y} = \phi_{00y} = 1)$. The expansion of the 95% interval for OR_{TY} to [0.37, 3.4] further suggests that the data add little information about OR_{TY} beyond that conveyed by the priors.

The simple simulation analysis just described can be extended to allow for selection bias and confounder adjustment (both measured and unmeasured) as well as for misclassification, although it can become quite unwieldy due the number of bias parameters and data cells involved (Greenland 2005). Unfortunately, it does not extend easily to situations in which all parameters (not just bias parameters) are given priors, whereas Bayesian computation has no such difficulty.

19.4 Pointwise Versus Probabilistic Analysis

While basic sensitivity or bias analyses are widely accepted and even promoted, there are difficulties if not objections to arbitrary elements in the priors are needed for probabilistic extensions (such as prior shape, e.g., normal vs. beta vs. trapezoidal). These difficulties are understandable given the lack of neutral and accepted conventions for these elements.

To address these elements, the analysis may be repeated using different identifying distributions and the results from different choices tabulated and contrasted. This process is called *prior sensitivity analysis*. Sensitivity analyses that instead directly vary bias parameters (*pointwise analyses* or deterministic analyses) are the special case of prior sensitivity analysis in which the varied priors are limited to point-mass priors (which assign 100% probability to just one value for the parameter).

From this viewpoint, pointwise analyses are of limited value compared to full probabilistic sensitivity analysis. Their main limitation is that they provide no explicit accounting for uncertainty about the bias parameters (Greenland 1998, 2005; Gustafson et al. 2001). The point-mass priors implicit in these analyses have no scientific justification given the uncertainty inherent in the parameters and do not escape the arbitrariness problem because they are based on a range of bias-parameter variation that is itself arbitrary. Too broad a range will lead to inclusion of values that would be recognized by subject experts as misleadingly absurd (e.g., a smoking prevalence of 90% in a general population), while too narrow a range will exclude important possibilities that cannot be ruled out. It is thus worthwhile to use genuine background information to create credible or plausible priors, or at least to specify the values used in a pointwise analysis.

In accounting for background information, there is a need to avoid giving that information more weight than it is worth, in recognition of the generalization problems touched on earlier. In particular, prior distributions should not be too narrow (incautious), lest our results remain overconfident. Of special note is that confidence intervals from other studies do not account for generalization problems and thus are too narrow to be used directly as prior intervals; at the very least, their probability percentage would need to be reduced below the original confidence percentage. For example, a 95% confidence interval for an effect from one study (even a perfect trial) should not be accorded 95% probability of containing the effect in another study, since generalization should reduce our confidence that the interval contains the latter effect.

Even with attention to such details, what appear to be cautious priors to one analyst may appear incautious to another. An ideal solution would allow each analyst to plug in their own priors and repeat the procedure, thus making the audience an active participant in sensitivity analysis. Macros and spreadsheets allowing such a solution have been offered (Steenland and Greenland 2004; Fox et al. 2005; Orsini et al. 2008; Lash et al. 2009); in principle these do not require audience access to the data, so confidentiality can be assured. Similar extensions to model-sensitivity analysis and influence analysis are also possible (e.g., by re-specifying external constraints on inclusion and exclusion criteria).

19.5 Some Theoretical Considerations for General Bias Analysis

This section concerns general formulations that come into play in analyzing more complex data and in multiple-bias modeling (Greenland 2005, 2009b; Vansteelandt et al. 2006; Molitor et al. 2009). It presumes some familiarity with mathematical statistics and may be skipped by less theoretical readers.

For the purposes of basic bias analysis of a single data set, the following definitions can be useful. A parameter in a data-generating model is *fully identified* by a system of estimating equations for the model parameters if the parameter is constant over the set of solutions to the system. More generally, a function of the model parameters is *fully identified* by the equations for the model parameters if the function maps the solution set to a unique value. At the opposite extreme, the function is *not identified* by the equations if it maps the solutions onto its entire range. Thus, in the confounding example (Table 19.3), the target parameter is not identified by its estimating equation $OR_{XY|U} \approx \exp(\widehat{\beta}_{XY})/\text{Bias}_U$, because given XY data with a finite $\widehat{\beta}_{XY}$ (and no further constraint), there is a solution for every possible value r of $OR_{XY|U}$; one merely chooses the bias parameters to make $\text{Bias}_U = \exp(\widehat{\beta}_{XY})/r$.

Between the extremes of full and no identification, a function is *partially identified* by the equations if it maps solutions to a proper subset of its range but not to a unique value. In bias settings this usually means the solutions are mapped into a subset of the range of reduced but positive dimension, although it may also or instead mean that the image of the solutions is sharply bounded inside the function range. The entire model is *non-identified* if some function of its parameters is only partially identified. An *identification problem* exists if the target parameter is not fully identified, for that means multiple possible values for that parameter arise from solutions to the estimating equations.

These are data-dependent definitions and so oriented toward likelihood and Bayesian analysis; they are modified for asymptotic sampling theory since non-identification whose probability goes to zero with increasing sample size can be ignored in that theory. Regardless, identification problems are traditionally solved by introducing artificial, arbitrary constraints (such as no bias for the target parameter) that at most have justification in a very narrow range of real circumstances (e.g., randomized trials with perfect compliance and no loss) and may be generalized only in very limited ways (e.g., to generalized linear semi-parametric models like Cox proportional hazards regression).

A common example is the assumption of *no residual selection bias or confounding* (often referred to as "no uncontrolled selection bias or confounding" although the terminology varies considerably). This condition corresponds to an ignorability condition that selection for analysis and treatment assignment are randomized conditional on the regression covariates, making selection and assignment independent of the potential outcomes that compose the target effect parameter (the ignorability being *strong* or *weak* according to whether it refers to joint or component-wise independence of potential outcomes and treatment assignment). Coupled with the assumption that the regression model is correct, the regression model for Y then coincides with the structural equation for Y (the Y-potential-outcome function), and further confounding concerns are obviated; unbiased conditional-effect estimation then reduces to the familiar problem of unbiased estimation of the regression of Y on treatment and the covariates. Alternatively, assuming instead correct models for selection and treatment-assignment probabilities, the fitted models for these probabilities can be used to estimate marginal effects (as in inverse-probability-weighting methods).

A problem with this traditional approach (of forcing identification through no-residual bias assumptions) is that it generates highly overconfident inferences: Confidence intervals so derived fail to cover the target near the nominal rate, and posterior intervals are far too narrow given the actual amount of external information available. This overconfidence follows from forcing unknown bias parameters to equal unique values, which takes no account of the huge uncertainty surrounding those parameters.

To reflect this uncertainty properly, Leamer (1974, 1978) and many others recommended instead one begin with non-identified models, and then examine the impact of augmenting the estimating functions with contextually based parameter constraints (Graham 2000; Gustafson et al. 2001; Greenland 2003, 2005, 2009a, b; Gustafson 2003, 2005). Those constraints are in turn derived from a *prior distribution* or *penalty function* for some or all of the parameters. If one allows for the fallibility of these contextual inputs, bias analysis then serves chiefly as an antidote or diagnostic to counterbalance the overconfident results from the conditionally randomized model. The conditionally randomized model remains a simple if doubtful reference point, arguably a necessary convention given the enormous number of prior distributions that would identify the target parameter.

Randomization assumptions correspond to prior distributions that assign positive probability to only a proper subspace of the non-identified parameters (making the priors degenerate with respect to the bias-parameter space). In the selection-bias example, random sampling corresponds to no association of X with S, which occurs in the subspace defined by $\pi_{s1y} = \pi_{s0y}$; in the confounding example, randomization induces no association of U with X, which occurs in the subspace defined by $\pi_{u1y} = \pi_{u0y}$. More generally, no net bias ($\text{Bias}_{S=1} * \text{Bias}_U * \text{Bias}_M = 1$ in the example) constrains all bias parameters to a manifold of much lower dimension than the entire bias-parameter space. Realistic priors rarely entail any such dimension reduction, even when they concentrate near lower-dimensional manifolds, and thus reveal the source of overconfidence from traditional identifying assumptions.

19.6 Conclusions

The present chapter has addressed settings in which a "correct" model for the data generation can never be known or approximated, and hence the data cannot tell us whether we are close to or far from the target, even probabilistically. In these settings, we cannot guarantee that inferences from a posterior distribution will be superior to inferences from our prior alone (Neath and Samaniego 1997). Thus, the importance of specific models and priors is de-emphasized in favor of providing a framework for sensitivity analysis across plausible models and priors. This framework need not be all-encompassing, because often just a few plausible specifications can usefully illustrate the illusory nature of an apparently conclusive conventional analysis.

When inference is mandated in these settings (as in pooled analyses to advise policy), we must admit we can only propose models that incorporate or are at least consistent with facts as we know them and that all inferences are completely dependent on these modeling choices (including non-parametric or semi-parametric inferences). There will be an infinite number of such models, and they will not all yield similar inferences, necessitating some sort of sensitivity analysis to provide a picture of the problem unless the effects in question are so large as to be obvious (as is typically the case in disease outbreaks).

In this task statistical modeling provides only inferential possibilities rather than inferences. Any analysis should thus be viewed as part of a sensitivity analysis which depends on external plausibility considerations to reach conclusions (Greenland 2005, 2009b; Vansteelandt et al. 2006). Results from single models are merely examples of what might be plausibly inferred, although just one plausible inference may suffice to demonstrate inherent limitations of the data.

Whatever their value for summarization, conventional models treat unknown parameters as if known and thus do not satisfy plausibility considerations. These unknowns include many bias parameters that can be forced to their null by successful design strategies but are probably not null in most observational settings. The use of prior distributions can provide more plausible inferences by allowing expansion of conventional models to include bias parameters as unknowns.

In any given observational context, progress beyond prior-based analyses can be made only by obtaining data from a study design that pins down a previously unknown bias parameter. Examples of such designs include studies with successful randomization, highly accurate measurement, or that have isolated a controversial effect in a setting where the effect is so large as to exceed plausible bias magnitudes. Because such designs are often infeasible (e.g., in the electromagnetic field-childhood cancer controversy), proper construction and use of prior distributions should become a component of statistical training for observational research.

Despite these considerations, there is no basis for mandating a bias analysis of every study or even most studies. For example, bias analysis is superfluous when conventional intervals show that no useful conclusion could be drawn from the study even if it were perfect apart from random error. More generally, rather than providing a bias analysis, a study may provide greater service by refraining from inference; instead it can focus on carefully reporting its design, conduct, and data in great detail to facilitate pooling and meta-analysis (Greenland et al. 2004). Inferences are best based on a more complete account of evidence than can be provided in a single study report, and thus the effort of bias analysis is more justifiable in research synthesis (Turner et al. 2009; Welton et al. 2009). Even there, bias analysis becomes essential only when doing risk assessment or when authors claim to offer near-definitive conclusions.

References

Breslow NE, Day NE (1980) Statistical methods in cancer research. Vol I: the analysis of case-control data. IARC, Lyon

Bross IDJ (1966) Spurious effects from an extraneous variable. J Chronic Dis 19:637–647

Bross IDJ (1967) Pertinency of an extraneous variable. J Chronic Dis 20:487–495

Brumback BA, Hernan MA, Haneuse S, Robins JM (2004) Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. Stat Med 23: 749–767

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu C (2006) Measurement error in nonlinear models, 2nd edn. Chapman and Hall, Boca Raton

Cole SR, Chu H, Greenland S (2006) Multiple-imputation for measurement-error correction (with comment). Int J Epidemiol 35:1074–1082

Copas JB, Li HG (1997) Inference for non-random samples (with discussion). J R Stat Soc Ser B 59:55–77

Cornfield J, Haenszel WH, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. J Natl Cancer Inst 22:173–203

Chu H, Wang Z, Cole SR, Greenland S (2006) Sensitivity analysis of misclassification: a graphical and a Bayesian approach. Ann Epidemiol 16:834–841

Eddy DM, Hasselblad V, Schachter R (1992) Meta-analysis by the confidence profile method. Academic press, New York

Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. Chapman and Hall, New York

Flanders WD, Khoury MJ (1990) Indirect assessment of confounding: graphic description and limits on effect of adjusting for covariates. Epidemiology 1:199–246

Fox MP, Lash TL, Greenland S (2005) A method to automate probabilistic sensitivity analyses of misclassified binary variables. Int J Epidemiol 34:1370–1376

Gail MH, Wacholder S, Lubin JH (1988) Indirect corrections for confounding under multiplicative and additive risk models. Am J Ind Med 13:119–130

Glymour MM, Greenland S (2008) Causal diagrams. Chapter 12. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 3rd edn. Lippincott-Williams-Wilkins, Philadelphia, pp 183–209

Graham P (2000) Bayesian inference for a generalized population attributable fraction. Stat Med 19:937–956

Greenland S (1998) The sensitivity of a sensitivity analysis. In: 1997 Proceedings of the biometrics section. American Statistical Association, Alexandria, pp 19–21

Greenland S (2001) Sensitivity analysis, Monte-Carlo risk analysis, and Bayesian uncertainty assessment. Risk Anal 21:579–583

Greenland S (2003) The impact of prior distributions for uncontrolled confounding and response bias. J Am Stat Assoc 98:47–54

Greenland S (2005) Multiple-bias modeling for observational studies (with discussion). J R Stat Soc Ser A 168:267–308

Greenland S (2008) Introduction to Bayesian statistics. Chapter 18. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 3rd edn. Lippincott-Williams-Wilkins, Philadelphia, pp 328–344

Greenland S (2009a) Bayesian perspectives for epidemiologic research. III. Bias analysis via missing-data methods. Int J Epidemiol 38:1662–1673. corrigendum (2010) Int J Epidemiol 39:1116

Greenland S (2009b) Relaxation penalties and priors for plausible modeling of nonidentified bias sources. Stat Sci 24:195–210

Greenland S, Lash TL (2008) Bias analysis. Chapter 19. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 3rd edn. Lippincott-Williams-Wilkins, Philadelphia, pp 345–380

Greenland S, Maldonado G (1994) The interpretation of multiplicative model parameters as standardized parameters. Stat Med 13:989–999

Greenland S, Mickey RM (1988) Closed form and dually consistent methods for inference on strict collapsibility in 2x2xK and 2xJxK tables. Appl Stat 37:335–343

Greenland S, Gago-Domiguez M, Castellao JE (2004) The value of risk-factor ("black-box") epidemiology (with discussion). Epidemiology 15:519–535

Gustafson P (2003) Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments. Chapman and Hall/CRC, Boca Raton

Gustafson P (2005) On model expansion, model contraction, identifiability, and prior information (with discussion). Stat Sci 20:111–140

Gustafson P, Le ND, Saskin R (2001) Case–control analysis with partial knowledge of exposure misclassification probabilities. Biometrics 57:598– 609

Gustafson P, McCandless LC, Levy AR, Richardson SR (2010) Simplified Bayesian sensitivity analysis for mismeasured and unobserved confounders. Biometrics 66:1129–1137

Hatch EE, Kleinerman RA, Linet MS, Tarone RE, Kaune WT, Auvinen A, Baris D, Robison LL, Wacholder S (2000) Do confounding or selection factors of residential wire codes and magnetic fields distort findings of electromagnetic fields studies? Epidemiology 11:189–198

Joseph L, Gyorkos TW, Coupal L (1995) Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol 141: 263–272

Kitagawa EM (1955) Components of a difference between two rates. J Am Stat Assoc 50: 1168–1194

Kraus JF, Greenland S, Bulterys MG (1989) Risk factors for sudden infant death syndrome in the U.S. Collaborative Perinatal Project. Int J Epidemiol 18:113–120

Lash TL, Fink AK (2003) Semi-automated sensitivity analysis to assess systematic errors in observational epidemiologic data. Epidemiology 14:451–458

Lash TL, Fox MP, Fink AK (2009) Applying quantitative bias analysis to epidemiologic data. Springer, New York

Leamer EE (1974) False models and post-data model construction. J Am Stat Assoc 69:122–131

Leamer EE (1978) Specification searches. Wiley, New York

Leamer EE (1985) Sensitivity analyses would help. Am Econ Rev 75:308–313

Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York

Lyles RH (2002) A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure. Biometrics 58:1034–1037

MacLehose RF, Gustafson P (2012) Is probabilistic bias analysis approximately Bayesian? Epidemiology 23:151–158

McCandless LC, Gustafson P, Levy AR (2007) Bayesian sensitivity analysis for unmeasured confounding in observational studies. Stat Med 26:2331–2347

McCandless LC, Gustafson P, Levy AR, Richardson SR (2012) Hierarchical priors for bias parameters in Bayesian sensitivity analysis for unmeasured confounding. Stat Med 31: 383–396

Molitor N-T, Best N, Jackson C, Richardson S (2009) Using Bayesian graphical models to model biases in observational studies and to combine multiple sources of data: application to low birth weight and water disinfection by-products. J R Stat Soc Ser A 172:615–638

Neath AA, Samaniego FJ (1997) On the efficacy of Bayesian inference for nonidentifiable models. Am Stat 51:225–232

Orsini N, Bellocco R, Bottai M, Wolk A, Greenland S (2008) A tool for deterministic and probabilistic sensitivity analysis of epidemiologic studies. Stata J 8:29–48

Pearl J (2009) Causality, 2nd edn. Cambridge University Press, New York

Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P (2004) Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol 159:882–890

Phillips CV (2003) Quantifying and reporting uncertainty from systematic errors. Epidemiology 14:459–466

Rosenbaum PR (2002) Observational studies, 2nd edn. Springer, New York

Saltelli A, Chan K, Scott EM (eds) (2000) Sensitivity analysis. Wiley, New York

Scharfstein DO, Daniels MJ, Robins JM (2003) Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes. Biostatistics 4:495–512

Schlesselman JJ (1978) Assessing effects of confounding variables. Am J Epidemiol 108:3–8

Schneeweiss S (2006) Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiol Drug Saf 15:291–303

Steenland K, Greenland S (2004) Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. Am J Epidemiol 160:384–392

Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG (2009) Bias modeling in evidence Synthesis. J R Stat Soc Ser A 172:21–47

VanderWeele TJ, Arah OA (2011) Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments and confounders. Epidemiology 22:42–52

Vansteelandt S, Goetghebeur E, Kenward MG, Molenberghs G (2006) Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. Stat Sin 16:953–980

Welton NJ, Ades AE, Carlin JB, Altman DG, Sterne JAC (2009) Models for potentially biased evidence in meta-analysis using empirically based priors. J R Stat Soc Ser A 172:119–136

Yanagawa T (1984) Case-control studies: assessing the effect of a confounding factor. Biometrika 71:191–194

Use of Health Registers

20

Reijo Sund, Mika Gissler, Timo Hakulinen, and Måns Rosén

Contents

R. Sund (✉)
Service Systems Research Unit, THL - National Institute for Health and Welfare, Helsinki, Finland

M. Gissler
Information Department, THL - National Institute for Health and Welfare, Helsinki, Finland

T. Hakulinen
Finnish Cancer Registry - Institute for Statistical and Epidemiological Cancer Research, Helsinki, Finland

M. Rosén
SBU – The Swedish Council on Technology, Assessment in Health Care, Stockholm, Sweden

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_5,

20.1 Introduction

Routine data are data collected continuously or at least repeatedly with some time intervals. They are collected in various ways, e.g., registration by the health services or by interviews with patients or population groups. The data could then be stored and administered in register format. Health registers contain data on persons with diseases or health-related events. The coverage can vary from a total registration to a population sample and from national to regional or local coverage. Data can be routinely collected for various reasons, from economic and administrative purposes to more strict epidemiological purposes.

We have limited our review to registers that allow individuals to be followed up. Such registers are routinely available in the Nordic countries, but also increasingly elsewhere. Most data in such registers contain event-based records, but registration of events with personal identification numbers or other identifiers allows observation of the health event sequences for each individual by using record linkage. Identifiers also make it possible to remove duplicate reports on the same events as well as add subsequent follow-up data later on. Anonymous data and statistical information have some value especially for descriptive epidemiological purposes, but the real value of register-based information systems depends on the possibility for record linkages.

This chapter is outlined in the following way. First, a presentation of registers will be made including types of registers, organization, contents, and variables in the registers as well as the quality of registers. Second, analytical options for register-based studies will be presented followed by examples of the usefulness and discussions on potentials and limitations of different designs. The ethical questions of privacy, autonomy, and confidentiality will also briefly be discussed.

20.2 Types of Registers

Most health registers can be classified into four basic categories: administrative registers, disease registers, reimbursement registers, and quality registers. Administrative registers record events that are needed for administrative, typically statistical purposes, and they cover, e.g., causes of death, births, hospital discharges, and population characteristics. Disease registers contain data related to some specific disorder, such as cancer or diabetes. Reimbursement registers compile monetary records such as health insurance claims, health-related social benefits, or reimbursable purchases of prescription drugs. Quality registers are collected in order to audit or to follow up treatment procedures and outcomes of care. An early example of quality register is the registry for total hip replacement surgery in Sweden which started in 1979 (Kärrholm 2010).

The most widespread and well-known registers are the causes of death and cancer registers. Most countries in the world have causes of death registers, mostly based on total registration, but in some cases on a sample of deaths, e.g., covering only deaths occurring in hospitals. Cancer registers are also frequently used around the world. In several countries, it is mandatory for the physician and other health care personnel to report all patients with diagnosed malignant neoplasms.

Medical birth registers are common in many countries, though not as widespread as the causes of death and cancer registers. All Nordic countries have introduced a separate medical birth register for detailed collection of parturients, deliveries, and newborns (Gissler et al. 1997; Gissler 2010). The registration covers both hospital births and planned or unplanned births outside hospitals. The follow-up extends usually until the child is discharged from the hospital or latest until the end of perinatal period, i.e., the first week of living. Information on deaths, however, can cover the whole infant period until 1 year of age. In addition, the register on congenital anomalies and birth defects is usually closely linked to the medical birth register, either as a separate register or as a part of the medical birth register.

Discharge registers are also common and typically contain data on inpatient hospitalizations with diagnoses and treatments of the patient. From an epidemiological point of view, these registers may be more difficult to interpret since traditionally they have captured only those patients who require inpatient admission. Currently, also outpatient hospital visits are increasingly recorded in many countries, but different care practices and strategies heavily affect whether patients with the same disease or medical condition are hospitalized or treated in outpatient settings in primary care.

Several countries have also introduced a national database to register prescription drugs (Furu et al. 2010). National prescription registers contain data on drugs dispensed at pharmacies to people receiving ambulatory care. Nordic countries include all prescribed and dispensed medication, excluding Finland, which registers only medication covered by the national health insurance system (currently 99% of prescribed medications). Over-the-counter (OTC) drugs and drugs dispensed elsewhere than in pharmacists, e.g., hospitals and outpatient clinics, are excluded in all prescription registers. The largest limitation with the Nordic prescription registers is the lack of information on the indication or diagnosis for the use of medicine. Finally, the data are based on purchases of medication, and there is no information if the medicine was actually taken or not. Also, information on the use of old drugs, other people's drug, or use of drugs purchased in other countries remains unregistered.

Besides these main types of health registers, there are also many local research registers covering other disease groups like cardiovascular and psychiatric diseases. An increasing phenomenon is the development of quality registers. In Sweden, there are up to 100 national quality registers covering treatment procedures and outcomes for different disease groups or medical interventions (Swedish Local Authorities and Regions 2011). New data sets or registers have also been created by record linkages, the most common ones being linkages between the population censuses and other available health registers to gather information on special disease, such as diabetes (Sund and Koski 2009; Carstensen et al. 2011). Registers can also be used to follow a cohort for a long time, and these ad hoc registers are administered by research groups. But these research registers cannot usually be considered as routinely collected or permanent registers, as the national health registers kept by the authorities. The purposes and epidemiological applications for these linkages will be presented later on in this chapter.

Countries with lacking disease registers can use data from health insurance systems as a basis for epidemiological research. These data can cover routinely collected notifications on sickness absence of employees or health outcomes of the insured population. One problem is that insurance systems and insured populations usually vary in many aspects, which may create bias and limit their comparability. Validity may also be impaired due to the fact that these data are generated for administrative purposes. Diagnoses and treatments may be influenced by aspects related to payments and accounting. Another problem is that the insurance registers are not always accessible because of strict interpretation of confidentiality legislation and rules.

In the following, we have mainly restricted our presentation to the five most common registers, i.e., causes of death, cancer, birth, prescription, and hospital discharge registers.

20.3 Organization of Registries

According to the IEA dictionary of epidemiology (Last 1988), the term "register" is applied to the file of data that can be related to a population base, i.e., the actual document, while the registry is the system of ongoing registrations. The organization of registries and collection of data differ from country to country, but they have always some basic governmental or other public funding. In most countries, the registries are organized within a governmental body such as national statistical offices, national boards of health, or public health institutes. Many disease registers, such as cancer registries, have more or less close links to disease-related societies from where they also can get a part of their funding.

The Nordic countries have a long tradition of collecting data in health registers (Table 20.1). They employ epidemiological registers of high quality covering the whole population. The causes of death have been registered in Sweden since 1751 (computerized from 1952), while the oldest cancer register in Denmark dates back to 1943. The registers of congenital malformations were established in Finland and Sweden in 1963–1964 as an early warning system and as a direct response to the thalidomide catastrophe. Norway was the first to start a nationwide medical birth

Table 20.1 Starting years for computerized nationwide health register data in Nordic countries

Register	Denmark	Finland	Iceland	Norway	Sweden
Hospital discharges/inpatient care	1977	1967	2001	2007	1987
Outpatient hospital/ambulatory care	1995	1998	2005	2007	2001
Birth	1973	1987	1982	1967	1973
Cancer	1943	1953	1955	1952	1958
Prescription	1994	1994	2006	2004	2005
Causes of death	1970	1971	1981	1951	1952

register in 1967. It includes information on mothers and children, e.g., diagnoses, birth weight and height, operations as well as maternal tobacco and drug use during pregnancy. In Finland, all hospital discharges have been recorded since 1967 and outpatient care in public hospitals since 1998. In Denmark, inpatient care has been recorded since 1977 and ambulatory care since 1995. The National Hospital Discharge Register covers all publicly run inpatient care in Sweden from 1987, and data on hospital outpatient care has been collected from 2001. The value of these registers grows continuously as time passes.

The routine registers may not always be optimal for research purposes. If the register is run by a statistical authority, there may be strict limitations, as to whether or how much register personnel are allowed to use their time in scientific research. It has proven useful to dedicate scientifically qualified staff to run the register (Jensen and Whelan 1991), since these people have a research interest in the data. When the register data are good enough for scientific research, they normally also are guaranteed to be very good for routine statistical production. Moreover, the dedicated personnel are very useful in helping other researchers with the data use by knowing the strengths and weaknesses and the most relevant method issues. In many countries, e.g., in Finland, the existence of such personnel has been guaranteed, e.g., for cancer research (Finnish Cancer Registry 2011). The Finnish Cancer Registry is a research organization specialized in statistics and studies, making use of the nationwide cancer register in Finland. In Sweden, the National Board of Health and Welfare is responsible for several other health registers in addition to the cancer register (EpC 2003). These research organizations have a multidisciplinary scientific staff consisting of epidemiologists, statisticians, physicians, social scientists, computer scientists, etc.

It is necessary to have legislation on the registration. In the Nordic countries, the existence of the health registers and the way they are run with responsibilities, rights, and obligations is based on laws. It is also important to have secured funding and the core scientific and clerical staff on permanent funding. The knowledge required for register keeping and the scientific work cannot possibly be maintained on short-term project funding contracts only.

In countries with numerous registers, it may be difficult for researchers to find all the data and get access to them. Therefore, infrastructure supporting and promoting register-based research has been established: The National Center for Register-based Research in Denmark was established in 2000, the Finnish Information Center for Register Research (ReTki) in 2003, and the European Center for Register-Based Health-Related Population Research (ECREPH) in Denmark in 2008.

20.4 Personal Identification Number and Other Variables in the Registers

A necessity in registers with data on individuals is a unique identity, usually in forms of a personal identification number (PIN) or a social security number (SSN). If the same PIN system is used in all data sources, it is possible to make deterministic record linkages within and between data from different registers and other sources of

Table 20.2 Common variables in health data registers in the Nordic countries

Register	Variables
Causes of Death Register	PIN (personal identification number), age, sex, place of residence, date and place of death, underlying cause of death, nature of the injury, contributing causes of death, autopsy (clinical or forensic)
Cancer Register	Name, PIN, age, sex, place of residence, site of tumor, histological type, basis of diagnosis, date of diagnosis, reporting hospital and department, and reporting pathology/cytology department
Medical Birth Register	Mothers' and live-born children's PINs, maternal sociodemographic background, previous pregnancies and deliveries, maternal diagnoses, care and interventions during pregnancy and delivery, and information on newborn health, diagnoses, care and interventions
Hospital Discharge Register/Patient Register	PIN, sex, age, place of residence, date of admission, date of discharge, acute/planned admission, main and secondary diagnoses, external cause of injury and poisoning, surgical procedures, and hospital/department These registers are in some countries enlarged to also include outpatient visits at hospitals.
Prescription Register	PIN, age, sex, and place of residence, prescriber, prescribing and dispensing date of the drug, Anatomical Therapeutic Chemical classification (ATC) code, amount in defined daily doses (DDD), and pharmacy details

relevant data. Some countries have to rely only on names, birth dates, and addresses as the identifier which can cause some bias and creates practical problems for follow-up as only probabilistic linkages can be performed.

The use of PINs in Nordic systems has created large data banks that are invaluable to the research community, since the data do not have to be collected from scratch but can be linked with PINs. For example, an (fictional) example of a Finnish personal identity code (PIC) is 131052-308T, where six first digits correspond to date of birth (day, month, and two last digits of the year), separator shows the century (+ for 1800, – for 1900, and A for 2000), number consisting of three next numbers distinguish persons with the same date of birth from each other and is even for women and odd for men, and the last digit is check digit calculated by modulo-31 algorithm (PRC 2011). The Finnish PIC was introduced and appointed to all living residents during 1960s and has a smart and informative structure adopted from the Swedish version (Ludvigsson et al. 2009) but improved with the excellent check digit algorithm. Virtually, everyone knows their own date of birth which means that they also remember their PIC without extra efforts and check digit as well as odd/even sex rule help to accept only valid PICs during registration.

Other common variables usually collected in the Nordic health registers are sex, age, and residential area of the individual (Table 20.2). The cause of death register usually includes data on underlying and contributory causes of death, place and date of death, and basis of cause certification. For the cancer registers, data on tumors, date of diagnosis, histological type, reporting department, hospital and pathology/cytology department, etc., are available. Medical birth registers generally include data on sex, weight, length, size of head, analgesia, birth conditions and operations of the children, but also data on the mother's previous gestation, smoking habits, and, in Sweden, also drugs taken during pregnancy. Hospital discharge registers include data on main and secondary diagnoses, external cause of injury and poisoning, surgical procedures, date of admittance and discharge, length of stay, and hospital and clinical department. Except operations, there are in general few or no data on type of interventions during hospital stay.

20.5 The Quality of Data in Registers

Register data have not been collected primarily for certain specific research design or study questions, and the quality and validity of routinely collected data can hardly be as consistent as for those obtained in clinical trials where predefined criteria for data collection are used. In fact, the validity of the data obviously depends on the intended use (Sørensen et al. 1996).

Most of the numerous validity studies of health registers have evaluated the quality from the epidemiological perspective and in general detected relatively good completeness of registration and validity of variables in the registers. For example, cancer registration in the Nordic countries is based on compulsory reports from all physicians and all pathologists, in both public and private administration. One Swedish study from the 1980s estimated the deficit in cancer registration to be 4.5% and less than 2% when the diagnosis had been histologically verified (Mattsson 1984). A Finnish study showed good coverage for solid tumors, but a roughly 10% underregistration for benign neoplasms of the central nervous system, chronic lymphatic leukemia, and multiple myeloma (Teppo et al. 1994). Technical quality control procedures are usually applied by cancer registries, e.g., the computers are programmed to detect invalid codes, inconsistent combinations of codes, duplicate registrations, and illogical time sequences. Examples of inconsistent combination of codes are testis cancer in a female and distant metastasis associated with carcinoma in situ. This kind of quality control reduces some types of errors but cannot deal with missing data or false primary diagnoses. Some cancers will never be diagnosed during the lifetime of individuals but may occur on the death certificate after an autopsy. Thus, changes and differences in autopsy rates between time periods and regions may affect the validity of cancer registration. The increase in incidence of prostate cancer during the last decade is to a large extent due to extensive PSA tests rather than a real increase in incidence (Tretli et al. 1996; Walsh 2002).

The accuracy of diagnoses in the causes of death registers varies considerably depending on the age of the deceased, underlying and contributing diseases and depending on practice variations among physicians and coders. In general, fatal diseases where the deceased has been treated for some time before death have good accuracy of diagnoses, e.g., most cancers and ischemic heart disease. The autopsy rates are usually higher for younger people and accidents, which makes this group of diagnoses quite reliable. A general problem in many cause of death registers are, however, the declining autopsy proportions (Lindström et al. 1997). In Sweden, the overall autopsy proportion was about 41% in 1980 and about 13% in 2009. In addition, it is known that there are variations in classification procedures by country (Percy and Muir 1989), and also regional variations within a country should be interpreted with caution due to potential practice variations.

Most of the variables in the Nordic Medical Birth Registers have been reported to have good or at least satisfactory validity (Gissler et al. 1997), and a systematic reviews of validity studies of Finnish and Swedish inpatient registers concluded that validity is high for many but not all diagnoses (Ludvigsson et al. 2011; Sund 2012). For example, about 6% of those with a diagnosis of acute myocardial infarction in the Swedish inpatient register had a false positive diagnosis, while 3% were false negative, i.e., should have had an acute myocardial infarction as the main diagnosis but had some other diagnosis in the register (Rosén et al. 2000). The underreporting in Swedish inpatient register has been estimated to be less than 2% for somatic short-term care. In 2001, personal identification numbers were missing in 0.4% of number of stays, and the main diagnosis was missing in 0.9% of the stays reported from Swedish hospitals. Such biases can be considered minor in epidemiological studies that contain hundreds of thousands of cases.

If validity needs to be evaluated for other than basic epidemiological purposes, it may require a much more complicated approach. For example, if there is not any obvious definition for the good quality of variable, several substantially different values can be considered technically equally good, or the quality requirements may be conditional on other available information, it is typically a good idea to prepare a conceptual model mapping the available variables explicitly to their theoretical counterparts and use such a model as a framework for the actual validity analyses (Sund et al. 2007).

In addition to specific validation studies, a reasonable criterion of quality of a register is also the common use of its data in scientific studies as each properly performed study should have been critical in the use of register-based data. Nordic health registers seem to fulfil this criterion as there are, in addition to specific validation studies for each register, thousands of published scientific articles based on data from these registers.

20.6 Methodological Approaches to Register-Based Data Analysis

If register-based study is compared to a standard scientific inquiry, the most important difference is that register data are of secondary nature, i.e., collected for other purpose(s) than the specific research question in hand (Cartwright and Armknecht 1980). This changes the nature of traditional empirical research process so that measurement and related data collection cannot be tailored to meet the exact needs or best practices available for certain well-defined research question, but one needs to opportunistically use the available data. The limitations of secondary data are determined by the choices made while producing data, such as a decision to collect only easily available data and using fixed classifications in the data production

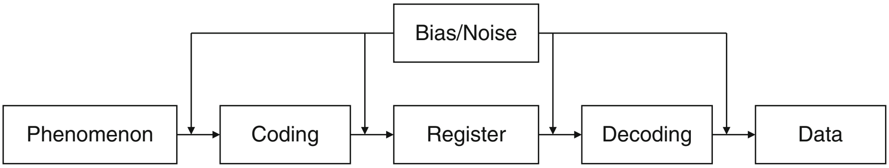


Fig. 20.1 Schematic diagram of information communication via administrative registers

with potentially varying local production practices. For example, imagine that your observational world is all that you see, hear, smell, taste, and feel. Would you trade your observational world for an old black-and-white photo when you could trade it for a high-quality three-dimensional digital movie with full color scale, including infrared (Sund 2003)?

Figure 20.1 is a schematic diagram illustrating information communication via registers. We may assume that some observable phenomenon exists, but it is impossible to make exact measurements for most phenomena. Some kind of simplified coding, however, can be used to describe most relevant things systematically. This measurement of phenomenon in terms of coded signal will then be stored in a database (or register). The noise and bias reflect the unavoidable measurement compromises, possible inconsistencies, or coding errors and coding practices existing in the stored signal. In order to utilize this signal (data in register), it must be decoded into suitable form for the current utilization purpose. This phase is also subject to noise and bias caused by incompatibility of choices and interpretations made by the data producer and the data user. Even the decoded signal (research data) is not a final phase in the research process, because further analysis and processing is needed in order to transform this data into actual information. Even though this is a very simple and technical representation of communication, it seems to capture the essential elements needed in the commonsense understanding of secondary data (Sund 2003).

It is particularly important to notice that it cannot be assumed that secondary data contain information, but information is something that has to be created from the data and pre-knowledge. This idea has been presented more formally in terms of infological equation

$$I = i(D, S, t),$$

which states that the information I is produced from the data D and the pre-knowledge S by the interpretation process i during time t (Langefors 1993). From this perspective, it becomes understandable that any sharing of data can only be a proxy for the sharing of information, because the unbiased communication would require the background knowledge S to be identical with the producer and the users of data. The problem can be dealt with by increasing the common pre-knowledge by offering descriptive data about data, i.e., metadata (Sundgren 1996). The key problem in using secondary register data in research is to find some shared perspective between the original and intended data utilization purposes in terms of available data and (more or less tacit) metadata (Sund 2007).

In practice, there are three options to utilize secondary data. First option is to use data that require no special pre-knowledge and are directly compatible with the current research problem. For example, data on dates of death are typically easily interpretable and can be used as such in most research problems. The second option is data abstraction, i.e., *enrichment* of data with special pre-knowledge (Shahar 1997). For example, there are no acute myocardial infarctions (AMIs) in hospital discharge or causes of death registers, but relevant information on them can be extracted by searching certain diagnoses from the registers by knowing that severe AMIs lead to hospital admission or death. Another example is a complication after surgical operation. That is a medical concept consisting of generalizations that apply across patients, but from individual-level hospital discharge data complications must be abstracted by using certain systematic rules, such as a list of particular diagnosis codes with appropriate time stamps recorded in the data. The third option is to slightly modify the original research problem to be more compatible with the available data. For example, it is difficult to identify reliably all hip fractures in risk population, needed for the calculation of hip fracture incidence, from the hospital discharge data, as there may be several readmissions and reoperations in addition to new fractures for each individual. In terms of data, the solution is to identify only the first hip fracture of each person. That is not the complete solution, because the standard incidence interpretation is not valid for such data, but also a justified epidemiological interpretation for the incidence of first hip fractures only is required (Sund 2007). Virtually, any use of register data requires data-sensitive preprocessing of data, and this preprocessing is typically the most important and time-consuming part of register-based data analysis.

Most data in health registers describe situation in the proximity of certain time stamped event such as cancer diagnosis, death, birth, hospital discharge, or drug purchase. Such observable events can be assembled to describe the history of an individual just as the basic idea of record linkage assumed (Dunn 1946). Such event history of an individual – event sequence – can also be interpreted as a sample path of a marked point process, which also makes it clear that a very general family of statistical hazard-rate models and other well-known methods in event-history framework suit well for the analysis of register-based data (Sund 2003). That framework contains also virtually all standard epidemiological methods. For example, health registers can be used to measure incidence, prevalence, mortality, and survival of different diseases over time and for different geographical areas and population groups. This is an important task for descriptive epidemiology (cf. see chapter ▸Descriptive Studies of this handbook), but the main advantage of registers for analytical purposes is that data from several sources can be linked together.

Record linkage of two or more national registers is one option. Another common application is that researchers use their own collected cohort as a baseline and then make follow-up in the national health data registers. The "Study of men born in 1913" is one such example of a cohort study where subjects with local registration of risk factors for cardiovascular disease were followed up for decades with health data registers (Tibblin et al. 1975). In addition to cohort studies, case-control as well as case-cohort studies are common study designs with register-based data. Also intergenerational studies have become feasible with register-based data as there is long history of data and typically population or birth registers contain identifiers for parent-child relationships. In fact, the possibility to link data allows creative construction of exposures and outcomes. For instance, long-term pregnancy outcomes of individuals who suffered major adverse events (such as the death of a child) can be followed (Li et al. 2009). Especially in Nordic countries with possibilities to deterministic record linkages and universal health systems, the health registers allow totally novel approaches to study many issues – only imagination is the limit. Some examples of studies are given in Sect. 20.8.

20.7 Potentials and Limitations of Register-Based Studies

In general, the advantages of using routine collected health registers as a main source for data are that they are already available, the number of observations is large, data are often nationwide and cover a whole population, and long follow-up periods as well as creative study designs become feasible via record linkages. By using large nationwide registers, on the one hand, a common critique of many studies to focus on men, on a limited age-group or on a specific geographical area can be avoided, and, on the other hand, it becomes feasible to study also rare outcomes. Registers also allow an assessment of outcomes in a real world situation that is not the case with randomized controlled trials with specific inclusion and exclusion criteria. Use of register data is also typically relatively cheap, especially compared to large surveys.

There are, of course, also shortcomings in national health registers. Some commonly mentioned problems are related to privacy and confidentiality issues, limited data availability, limited number of collected variables, poor population coverage, short registration period, limited record-linkage possibilities, lack of clinically relevant data, variation in data format, coding systems and coding practices, poor completeness of registration, limited accuracy of registered data, data processing issues, large size of data, possibility to discover chance occurrences, and changes in time in any of the above mentioned (Potvin and Champagne 1986; Connell et al. 1987; Bright et al. 1989; McDonald and Hui 1991). Many of the problems are, however, related to particular study designs or properties of certain known data sets, and therefore, their relevance should be evaluated separately for each study (Motheral et al. 2003). Quality of data can differ, but as described above, they are usually manageable for epidemiological research.

A more general problem is that the effective use of register-based data presumes skills in at least four fields: in computer science, statistics, the principles of measurement, and the theory of subject matter (David 1980). One can seldom be an expert on all of them, which makes collaboration the key to obtaining the best results.

Without doubt, accessibility to individual-based data in register research has a tremendous advantage when it comes to causal inferences. There are two fundamental pitfalls in this regard: Since all the data are typically available at the same time in register-based studies, it is easy to forget the direction in which time moves leading to temptation to organize the data analysis as prospective while conditioning on a future event. Another approach that is strictly against conservative causal reasoning is to "fish" or to "shop" interesting patterns from the data without any well-defined hypotheses.

In the following, some examples of the usefulness of register-based studies will be demonstrated, mainly when individual data are a necessity.

20.8 Examples of the Usefulness of Health Registers

Most of the Nordic registers are nationwide and cover the whole population. It may, however, be questioned, particularly in larger countries, whether it is not sufficient to do periodic surveys or geographically limited registration for the assessment of the disease occurrence in a country. The answer may depend on the disease or medical condition and on the desired goals of the registration or study. For rare outcomes, it is important to have large samples in order to come to conclusive answers. In environmental studies, it is an advantage to have data covering different areas. This will be illustrated later in this section. If epidemiological data are used for planning purposes like identifying needs of preventive actions, studies of equity in access to care or allocating health care resources, it is also important to have data from all parts of a country.

Thousands of scientific articles have been published based on data from national registers around the world and particularly from the Nordic countries. In many cases, the studies have added knowledge to the existing state of the art or later been replicated with other kinds of epidemiological studies, in others it has been nearly impossible for economic reasons to conduct studies without utilizing the existing registers. In general, register-based studies are very cost-effective due to the fact that data have already been collected for a long time period. Illustrations of the usefulness of register-based studies will be presented under headings: Environmental Epidemiology, Social Epidemiology, Pharmacoepidemiology, and Other Examples. For more examples in psychiatric epidemiology, see Miettunen et al. (2011).

20.8.1 Environmental Epidemiology

If the goal is to give answers related to environmental exposures, it is useful to have the registration in the whole country and over a longer period of time. For example, by using data from cancer register, it is possible to count the observed number of cancer cases in the area of interest and to compare them to expected numbers on the basis of suitable reference populations (e.g., neighbors or the cancer control region). It is also possible to evaluate the historical development in the area itself and in the other comparable areas. All these evaluations can be done really quickly when the historical database and the software are in an appropriate condition. The cancer register has been an important source, e.g., for analyzing the association between residential radon exposure and lung cancer in Sweden (Pershagen et al. 1994) and the potential effects of magnetic fields (Feychting and Ahlbom 1993).

Also the adverse effects of radioactive fallout following the Chernobyl accident in 1986 have been studied with registers. In Finland, the exposure level varied strongly between areas that were quite irregular in shape and did not coincide with health districts, but with data from the cancer registry, it was possible to reconstruct historical trends in childhood leukemia incidence in each of these different irregular areas and to study whether any changes in these trends could be related to the exposure level in the area. The answer was negative for the periods of 1986–1988 and 1989–1992 (Auvinen et al. 1994). The Swedish medical birth register was used to show that there was no significant adverse pregnancy outcome in Sweden after the Chernobyl accident (Ericson and Källén 1994).

More than 500,000 cleanup workers from the former Soviet Union were forced to clean the accident site in Chernobyl and its environment. The majority of the 500,000 cleanup workers came from different parts of the former Soviet Union where no reliable cancer registration existed (IPHECA 1996). For the almost 5,000 Estonian workers, however, it was possible to estimate whether the cancer risk of the workers had been affected, since cancer cases had been reliably registered in Estonia since 1978, and registration continued also after the accident and the return of the cleanup workers. The result has been negative for the period of 1986–1993 (Rahu et al. 1997). However, the former workers had an increased risk of dying through suicide, as shown by the causes of death register. Biological dosimetry done on the workers has also indicted that the range of doses the workers received was not likely to cause markedly increased cancer risks (Bigbee et al. 1997).

Cancer registration also revealed excess cancer risks related to the Chernobyl accident: children having lived in the nearby areas of Chernobyl experienced an epidemic of thyroid cancer (Kazakov et al. 1992). This was totally unexpected given the scientific knowledge available at the time of the accident.

Another example on routine register use is provided by a much smaller local exposure. In 1987, increased concentrations of chlorophenol were detected in the drinking water source of Järvelä, the center of Kärkölä municipality in Southern Finland (Lampi et al. 1992). According to the International Agency for Research on Cancer, exposure was related to possible increased risks at eight sites of cancer. The cancer registration in Finland covered Kärkölä, the other municipalities in the same local health care district and all the municipalities in the corresponding cancer control region. It was therefore very easy to quantify whether the population in the exposed municipality had faced any increased risks since 1953 related to the suspected eight forms of cancer or related to any other cancer. The results showed an increased risk of non-Hodgkin lymphoma and of soft-tissue tumors during the most recent 15 years, but not before that time. Subsequent case-control studies in the local health care district based on the material from the cancer registry confirmed that only these two cancers were concerned and the excess risk could be related to chlorophenol exposure. The exposure had come from a saw mill in Järvelä and was related to antifungal treatment of timber. Subsequent sediment analyses revealed that the exposure had been there for decades.

For a more detailed discussion of environmental epidemiology in general, we refer to chapter ▸Environmental Epidemiology of this handbook.

20.8.2 Social Epidemiology

Occupations at high risk for cancer and premature mortality have been followed and analyzed using record linkages of population censuses with the cancer register or the causes of death register. For example, up to 45 years of cancer incidence data by occupational category have been collected for the 15 million working aged persons in the 1960–1990 censuses in the Nordic countries (Pukkala et al. 2009). The study was undertaken as a cohort study with linkage of individual records based on the personal identity codes used in all the Nordic countries, and it was able to repeat most of the confirmed associations between occupations and cancers. Overall cancer risk was lowest in domestic assistants, seafarers, farmers, gardeners, and teachers, and highest in waiters, tobacco workers, seamen, and chimney sweepers. Almost all mesotheliomas are associated with asbestos exposure, and plumbers, seamen, and mechanics were the occupations with the highest risk to such neoplasm. Outdoor workers such as fishermen, gardeners, and farmers had the highest risk of lip cancer, while the lowest risk was found among indoor workers such as physicians and artistic workers. Increased nasal cancer risk, associated with exposure to wood dust, was found among woodworkers. Waiters and tobacco workers had a high risk of lung cancer, probably attributable to active and passive smoking. Miners and quarry workers also had a high lung cancer risk, which might be related to their exposure to silica dust and radon daughters. For more comprehensive view on occupational epidemiology see chapter ▸Occupational Epidemiology of this handbook.

The national registers have also been used continuously to study social inequalities, in healthcare and in mortality by record-linking health data registers with population censuses (Hallqvist et al. 1998; Manderbacka et al. 2009; Henriksson et al. 2010; Mortensen et al. 2011; Tarkiainen et al. 2012) and to examine the adverse health outcomes for vulnerable groups, e.g., psychiatric patients, immigrants, and single mothers (Ringbäck Weitoft et al. 1998, 2000; Wahlbeck et al. 2011).

Many studies of the health and social outcomes of single parents and their children have been conducted by linking several registers, i.e., population census, total enumeration income survey, hospital discharge register, and the causes of death register. The studies show increased premature mortality and morbidity both for single parents and their children, even after adjustments for socioeconomic factors and previous somatic and psychiatric inpatient history (Ringbäck Weitoft et al. 2000, 2003). Socioeconomic factors, especially a lack of economic resources, explained some of the disadvantages. Still, the results indicated an independent excess risk for single parents irrespective of socioeconomic factors and health selection into single parenthood. Similar results have been received in the long-term follow-up of women with alcohol or drug abuse during pregnancy and their children (Kahila et al. 2010; Sarkola et al. 2011).

Other aspects related to social epidemiology are discussed in chapter ▸Social Epidemiology of this handbook.

20.8.3 Pharmacoepidemiology

The study of vitamin K and childhood cancer well illustrates the advantages of large national health registers. A case-control study by Golding et al. (1992) indicated that intramuscular vitamin K administration doubled the risk of childhood cancer compared to oral administration. This result created much concern especially in Sweden, because intramuscular administration was recommended by the National Board of Health and Welfare. A study based on the medical birth and cancer registers was initiated immediately, and it showed no increased risk of childhood cancer (Ekelund et al. 1993). Later studies have confirmed the Swedish results. There were several differences between the British case-control study and the Swedish register-based study. One was sample size: the case-control study included 195 cases and 558 controls, while the register-based study included more than 2,300 childhood cancers and 1.3 million controls. In the register-based study, data were already available in the registers, and data from the medical birth and cancer register were merged by linking individual records. Supplemented with data on maternity hospital routines for vitamin K administration, the study was complete within a few months.

Another example is the possibly evaluated cancer risk associated with insulin glargine use. A large observational study suggested that use of insulin glargine is, after adjustment for dose, associated with a possible increase in tumor risk in humans (Hemkens et al. 2009). According to the editorial in the same issue, interpretation of the analysis proved controversial, but the implications were serious (Smith and Gale 2009). A special advisory group convened by the European Association for the Study of Diabetes agreed that the findings should not be published in isolation, and therefore three other large observational analyses were commissioned to investigate the safety of insulin glargine (Jonasson et al. 2009; Colhoun et al. 2009; Currie et al. 2009), and several other studies from many countries have been published since then. The evidence remains still somewhat inconclusive, but this case is an excellent example of how several countries could rapidly react and produce more information on the issue that would not have been feasible to study without good registers that already include the necessary data.

Since all the Nordic countries have initiated a prescription register between 1994 and 2006, Nordic studies on drug exposure have become feasible. For example, a study covering more than 1.6 million births showed that the use of selective serotonin-reuptake inhibitors (SSRI-antidepressants) in late pregnancy increased risks of persistent pulmonary hypertension of the newborn (Kieler et al. 2012).

A Finnish register-based study showed that long-term treatment with antipsychotic drugs of patients with schizophrenia was associated with lower mortality than no drug use and that clozapine seems to be associated with a substantially lower mortality than any other antipsychotics (Tiihonen et al. 2009). Studies have also shown inequity in access to drugs, e.g., immigrants from outside the European Union did not get all the recommended medication after myocardial infarction to the same extent as the Swedish citizens (Ringbäck Weitoft et al. 2008).

For more details on pharmacoepidemiological studies, refer to chapter ▶Pharmacoepidemiology of this handbook.

20.8.4 Other Examples

The medical birth registers have been used extensively in reproductive epidemiology, e.g., to analyze the short- and long-term risks of smoking during pregnancy (Ericson et al. 1991; Cnattingius and Haglund 1997; Ekblad et al. 2010), teenage pregnancy outcomes (Otterblad Olausson et al. 1999), and effects on children born after in vitro fertilization (Bergh et al. 1999; Klemetti et al. 2005). Since the registration of births has continued for decades in the Nordic countries, long-term follow-up (Paananen and Gissler 2011) and intergenerational studies (Nordtveit et al. 2008) have become feasible. As an example of Nordic collaboration is the study in which a cohort of 100,000 Nordic children born after in vitro fertilization treatment was built up to follow the children from conception until childhood and early adulthood (Henningsen et al. 2011).

A family-cancer database has been constructed in Sweden by linking population registers and the national cancer register in order to study familial cancer risks (Hemminki and Vaittinen 1998). The database includes approximately six million persons and more than 30,000 cancers in offspring diagnosed at ages 15–51 years and their parents. Numerous studies have been published based only on this database (Hemminki and Vaittinen 1998; Hemminki and Li 2001).

Survival analysis of cancer patients is a common application where registers have been used as extensive sources. Linking the date of cancer incidence in cancer registers with the date of death in the cause of death register create opportunities for survival analysis. Many large survival analyses have been published based on registers (Stenbeck and Rosén 1995; Dickman et al. 1999; Storm et al. 2010). The EUROCARE-4 study included follow-up of nearly 2.7 million cancer cases (Capocaccia et al. 2009). The social dimension of cancer survival has also been investigated in some studies by linking population census and cancer register (Vågerö and Persson 1987; Dickman et al. 1997). One methodological problem in survival analysis is that survival analysis by definition must be based on historical data, thus, assessing the effects of old treatment strategies. This problem is universal irrespective if the study is based on registers or not. However, one way to reduce this problem is to conduct period analysis using the latest information available (Brenner and Hakulinen 2002). Further methodological aspects are discussed in chapter ▶Survival Analysis of this handbook.

20.9 Quality Control in Medical Care Using Registers

There has also been a general trend worldwide to improve systems for effectiveness and quality control in medical care (OECD 2002). Outcomes of care have often been assessed by a mortality analysis. One famous example is the coronary bypass mortality study in New York State (Hannan et al. 1994, 1995) where they gathered clinical data whereby results could be adjusted for risks or patient mix. During the study period, mortality declined by 41% in the study area while it went down by only 18% in the rest of the country. The improvement was claimed to partly be due to the fact that less successful teams abandoned the market and partly to quality improvements by the other teams.

In Sweden, the Federation of Municipalities and County Councils (currently Swedish Local Authorities and Regions) and the National Board of Health and Welfare collaborate at the national level in providing financial and other kinds of support for creation and development of the national quality registers that routinely collect data on treatment procedures and outcomes of care (Garpenby and Carlsson 1994). Similar developments can be seen internationally, both in the Nordic and in other European countries. The hip fracture register in Sweden started in 1988 but expanded to a larger project where data are compared and analyzed between several European countries (Parker et al. 1998).

In Finland, a comprehensive performance monitoring system for several disease groups based entirely on existing nationwide registers has been constructed (Häkkinen 2011). A microeconomic disease-based strategy, in which the episodes of care and related costs are (re)constructed using record-linkable register data from several registers, has turned out to be a fruitful approach to assess the performance of health care systems (Peltola et al. 2011). In fact, the use of routinely collected registers is a very cost-effective way to produce useful information on performance, effectiveness, and quality of treatment without need for extra data collection. For example, it seems that some of the indicators that are typically calculated from the data collected with separate hip fracture audits can be derived directly from the routine Finnish registers (Sund 2008; Sund et al. 2011). Also this performance monitoring approach based on routine registers only has been extended to several European countries with individual-level linkable data (Häkkinen et al. 2011). Probably the best results could be obtained by combining the collection of separate quality registers (collection of important clinical data on selected disease events) and routine registers (common registration for all diseases).

20.10 Ethics, Confidentiality, and Legislation

The principles of autonomy, doing good, doing no harm, justice, and solidarity must guide decisions on how to administer national registers. The decision is a trade-off between benefits and risks (Allebeck 2002). The registers and their use are governed by national legislation.

There is one main difference between research based on health registers and other research projects where it is necessary to collect data on individuals. In the latter, there is a need for informed consent from all participants. This is not feasible for routinely collected national health registers. For these kinds of routinely collected national data banks, it would be practically and economically impossible to apply the informed-consent rule for data collection and for secondary use of register data in scientific research. To do so would substantially hamper clinical work and take resources from other important health service tasks. One may say that the national parliaments and governments in the Nordic countries have given informed consent on behalf of the population by national legislation. This exception from the rule of informed consent is based on the judgment that the benefits far outweigh the negative consequences. Here principles of doing good and justice or solidarity outweigh that of autonomy. This solution also follows the EU Directive on Data Protection (95/46), which states that Member States may, for reasons of substantial public interest, by national law or by decision of the supervisory authority, lay down exemption for the requirement for informed consent of each registered person (European Union 1995).

Another important topic is the risk of violating individual integrity. This risk of doing harm may be twofold: the risk of unlawful trespass/encroachment of data on individual diseases and medical conditions and the perceived uneasiness/discomfort at just being registered. No system could guarantee 100% security, but after more than five decades of administering health data registers in the Nordic countries, there is no known case of misuse or data leakage to unauthorized persons. Also the register keepers do their best in preventing any threat to data protection. In most cases, register-based studies can be performed by using anonymous or pseudonymized data sets without direct identifiers. The remote use of ad hoc research data located on the server of the register keeping organization prevents the problems related to delivering the data set with discs or memory sticks.

That some people feel discomfort at just being registered is a negative aspect we must consider seriously. Public confidence in health data registers is influenced by mass media debate and knowledge of how the registers are being handled. This confidence could vary from country to country. The Directorate-General Justice, Freedom, and Security at European Union regularly studies citizens' opinions on data protection. In the Eurobarometer survey in 2008 (Eurostat 2008), two out of three respondents in the European Union (64%) were very or fairly concerned on data protection issues. In general, older people and more educated were the most concerned population groups. By country, Finland had one of the lowest percentages (35%), while Sweden (75%) and Denmark (72%) were countries with relatively high proportion of people with data protection concerns. Also the percentage of citizens who are very concerned on data protection varies substantially between Sweden and Denmark (45–46%) and Finland (less than 5%). Dissemination of the purposes and the usefulness, and the careful administration, of these registers are therefore important and never-ending responsibilities for administrators and users.

It may well be concluded that it is worth having routine registers and registries also for giving a good basis for scientific research. Thus, scientific research should not primarily be seen as a secondary user of registers created mainly for other purposes. The only justification of registration is that the registered data are used. The use has to be guaranteed by securing the manpower and resources and by finding a correct balance between the individual's right to privacy and its protection and the right of the individual and the mankind to benefit from research knowledge based on data registers (International Association for Cancer Registries 1992).

For a broad discussion of ethical aspects in epidemiological research, refer to see chapter ▶Ethical Aspects of Epidemiological Research of this handbook.

20.11 Conclusions

The short summary of conducted studies presented in Sect. 20.8 shows clearly both the present and the future benefits of register-based epidemiological research. A small selection of studies using routine data and health registers presented here could also have been conducted by collecting new data sets. In that case, however, one would have to accept the use of much greater resources and more time before answers to the research questions were available. In some cases, it is not even feasible to conduct a study without national registers.

So far, the nationwide health registers with data linkable at individual level have mainly been a privilege of the Nordic countries. Smart choices to start routine collection of data and to introduce personal identity codes during the 1950s and 1960s have resulted in an invaluable gold mine for health research and in a laboratory environment for developing methodology for register-based data analysis. From this perspective, one may question why the utilization has not been as intensive as it could have been, especially after the argument that it would be unethical not to use available data to improve health in the population. A simple answer would be to state that the resources have been limited and should not be used for data with unknown quality. This may be partly true at least in terms of attitudes.

It is, of course, crucial that the quality of data in the health registers is good enough for research purposes. Many validity studies of the registers have been conducted indicating mostly good data quality and reliable results provided that the data are analyzed with care. In fact, the main challenge in the register-based research is the fact that there is not any simple solution to transform raw secondary register data into useful information, but a wide variety of expertise is required for innovative results. However, regardless of such pragmatic problems, only the imagination defines the limits of register-based data analyses, and every new study conducted with register-based data provides evidence on the usefulness of the registers.

In the future, the amount of register data will increase. For example, hospital discharge registers have been transformed to hospital patient registers where all visits to hospitals are included, and even information on primary health care visits are to be collected to central registers. As also the electronic health records can be seen a kind of health register, there will be more countries with register data and closer correspondence between clinical work and register-based surveillance systems in the future.

In the Nordic countries, the value of health registers increases all the time as longer and longer follow-up times and intergenerational studies become feasible. As the examples in Sect. 20.8 showed, there have been collaboration projects in which data from several Nordic countries have been pooled to provide larger data sets. This kind of collaboration is likely to be increasing in the future. It will also be essential to confirm that other data, such as biobanks or health surveys, will remain linkable to health register data as that will open novel perspectives for many issues.

References

Allebeck P (2002) The revised Helsinki declaration: good for patients? Good for public health? Scand J Public Health 30:1–4

Auvinen A, Hakama M, Arvela H, Hakulinen T, Rahola T, Suomela M, Söderman B, Rytömaa T (1994) Fallout from Chernobyl and incidence of childhood leukemia in Finland, 1976–92. BMJ 309:151–154

Bergh T, Ericson A, Hillensjö T, Nygren K-G, Wennerholm UB (1999) Deliveries and children born after in-vitro fertilization in Sweden 1982–1995: a retrospective cohort study. Lancet 354:1579–1585

Bigbee WL, Jensen RH, Veidebaum T, Tekkel M, Rahu M, Stengrevics A, Auvinen A, Hakulinen T, Servomaa K, Rytömaa T, Obrams GI, Boice JD Jr (1997) Biodosimetry of Chernobyl cleanup workers from Estonia and Latvia using the glycophorin A in vivo somatic cell mutation assay. Radiat Res 147:215–224

Brenner H, Hakulinen T (2002) Up-to-date long-term survival curves of patients with cancer by period analysis. J Clin Oncol 20:826–832

Bright RA, Avorn J, Everitt DE (1989) Medicaid data as a source for epidemiological studies: strengths and limitations. J Clin Epidemiol 42:937–945

Capocaccia R, Gavin A, Hakulinen T, Lutz JM, Sant M (eds) (2009) Survival of cancer patients in Europe, 1995–2002: The EUROCARE 4 study. Eur J Cancer 45:901–1094

Carstensen B, Kristensen JK, Marcussen MM, Borch-Johnsen K (2011) The national diabetes register. Scand J Public Health 39(Suppl 7):58–61

Cartwright DW, Armknecht PA (1980) Statistical uses of administrative records. Rev Public Data Use 8:13–27

Cnattingius S, Haglund B (1997) Decreasing smoking prevalence during pregnancy in Sweden: the effect for small-for-gestational-age birth. Am J Public Health 87:410–413

Colhoun HM, SDRN Epidemiology Group (2009) Use of insulin glargine and cancer incidence in Scotland: a study from the Scottish Diabetes Research Network Epidemiology Group. Diabetologia 52:1755–1765

Connell FA, Diehr P, Hart LG (1987) The use of large data bases in health care studies. Annu Rev Public Health 8:51–74

Currie CJ, Poole CD, Gale EA (2009) The influence of glucose-lowering therapies on cancer risk in type 2 diabetes. Diabetologia 52:1766–1777

David M (1980) Access to data: the frustration and utopia of the researchers. Rev Public Data use 8:327–337

Dickman PW, Gibberd RW, Hakulinen T (1997) Estimating potential savings in cancer deaths by eliminating regional and social class variation in cancer survival in the Nordic countries. J Epidemiol Community Health 51:289–298

Dickman PW, Hakulinen T, Luostarinen T, Pukkala E, Sankila R, Söderman B, Teppo L (1999) Survival of cancer patients in Finland 1955–1994. Acta Oncol 38(suppl. 12):1–103

Dunn HL (1946) Record linkage. Am J Public Health 36:1412–1416

Ekblad M, Gissler M, Lehtonen L, Korkeila J (2010) Prenatal smoking exposure and the risk of psychiatric morbidity into young adulthood. Arch Gen Psychiatry 67:841–849

Ekelund H, Finnström O, Gunnarskog J, Källén B, Larsson Y (1993) Administration of vitamin K to newborn infants and childhood cancer. BMJ 307:89–91

EpC, Centre for Epidemiology, National Board of Health and Welfare (2003) A finger on the pulse. Monitoring public health and social conditions in Sweden 1992–2002. EpC, Stockholm

Ericson A, Källén B (1994) Pregnancy outcome in Sweden after the Chernobyl accident. Environ Res 67:149–159

Ericson A, Gunnarskog J, Källén B, Otterblad Olausson P (1991) Surveillance of smoking during pregnancy in Sweden, 1983–1987. Acta Obstet Gynaecol Scand 70:111–117

European Union (1995) Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML. Accessed 8 Jan 2013

Eurostat. Flash Eurobarometer 225/2008. Data protection in the European Union. Citizens' perceptions. Analytical report. http://ec.europa.eu/public_opinion/flash/fl_225_en.pdf. Accessed 8 Jan 2013

Feychting M, Ahlbom A (1993) Magnetic fields and cancer in children residing near Swedish high voltage power lines. Am J Epidemiol 138:467–481

Finnish Cancer Registry – Institute for Statistical and Epidemiological Cancer Research (2011) Cancer incidence in Finland 2008 and 2009. Cancer Society of Finland Publication No. 84. Finnish Cancer Registry, Helsinki

Furu K, Wettermark B, Andersen M, Martikainen JE, Almarsdottir AB, Sørensen HT (2010) The Nordic countries as a cohort for pharmacoepidemiological research. Basic Clin Pharmacol Toxicol 106:86–94

Garpenby P, Carlsson P (1994) The role of national quality registers in the Swedish health service. Health Policy 29(3):183–195

Gissler M (2010) Registration of births and induced abortions in the Nordic countries. In: Finnish Demographic Yearbook XLV 2010. The Population Research Institute. Vammalan Kirjapaino Oy, Vammala, pp 171–178

Gissler M, Louhiala P, Hemminki E (1997) Nordic medical birth registers in epidemiological research. Eur J Epidemiol 13:169–175

Golding J, Greenwood R, Birmingham K, Mott M (1992) Childhood cancer, intramuscular vitamin K, and pethidine given during labour. BMJ 305:341–346

Häkkinen U (2011) The PERFECT project: measuring performance of health care episodes. Ann Med 43(Suppl 1):S1–S3

Häkkinen U, Malmivaara A, Sund R (2011) PERFECT-conclusions and future developments. Ann Med 43(Suppl 1):S54–S57

Hallqvist J, Lundberg M, Diderichsen F, Ahlbom A (1998) Socioeconomic differences in risk of myocardial infarction 1971–1994 in Sweden: time trends, relative risks and population attributable risks. Int J Epidemiol 27:410–415

Hannan EL, Kilburn JF, Racz M, Shields E, Chassin MR (1994) Improving the outcomes of coronary artery bypass surgery in New York State. JAMA 271:761–766

Hannan EL, Siu AL, Kumar D, Kilburn H Jr, Chassin MR (1995) The decline in coronary bypass surgery mortality in New York State. The role of surgeon volume. JAMA 273:209–213

Hemkens LG, Grouven U, Bender R, Günster C, Gutschmidt S, Selke GW, Sawicki PT (2009) Risk of malignancies in patients with diabetes treated with human insulin or insulin analogues: a cohort study. Diabetologia 52:1732–1744

Hemminki K, Li X (2001) Familial colorectal adenocarcinoma from the Swedish Family-Cancer database. Int J Cancer 94:743–748

Hemminki K, Vaittinen P (1998) National database of familial cancer in Sweden. Genet Epidemiol 15:225–236

Henningsen A-K, Romundstad L B, Gissler M, Nygren K-G, Lidegaard Ø, Skjaerven R, Tiitinen A, Nyboe Andersen A, Wennerholm U-B, Pinborg A (2011) Infant and maternal health monitoring using a combined Nordic database on ART and safety. Acta Obstet Gynecol Scand 90:683–691

Henriksson G, Weitoft GR, Allebeck P (2010) Associations between income inequality at municipality level and health depend on context – a multilevel analysis on myocardial infarction in Sweden. Soc Sci Med 71:1141–1149

International Association for Cancer Registries (1992) Guidelines on Confidentiality in the Cancer Registry. International Agency for Research on Cancer IARC Internal Report No. 92/003, Lyon

IPHECA (International Programme on the Health Effects of the Chernobyl Accident) (1996) Health consequences of the Chernobyl accident. Results of the IPHECA pilot projects and related national programmes. World Health Organization, Geneva

Jensen OM, Whelan S (1991) Chapter 4: planning a cancer registry. In: Jensen OM, Parkin DM, Maclennan R, Muir CS, Skeet RG (eds) Cancer registration: principles and methods. IARC Scientific Publications No. 95. IARC, Lyon, pp 22–28

Jonasson JM, Ljung R, Talbäck M, Haglund B, Gudbjörnsdòttir S, Steineck G (2009) Insulin glargine use and short-term incidence of malignancies-a population-based follow-up study in Sweden. Diabetologia 52:1745–1754

Kahila H, Gissler M, Sarkola T, Autti-Rämö I, Halmesmäki E (2010) Maternal welfare, morbidity and mortality 6–15 years after a pregnancy complicated by alcohol and substance abuse: a register-based case-control follow-up study of 524 women. Drug Alcohol Depend 111: 215–221

Kärrholm J (2010) The Swedish Hip Arthroplasty Register (www.shpr.se). Acta Orthop 81:3–4

Kazakov VS, Demidchik EP, Astakhova LN (1992) Thyroid cancer after Chernobyl. Nature 359:21

Kieler H, Artama M, Engeland A, Ericsson Ö, Furu K, Gissler M, Nørgaard M, Beck Nielsen R, Stephansson O, Valdimarsdóttir U, Zoega H, Haglund B (2012) Selective serotonin-reuptake inhibitors during pregnancy and risks of persistent pulmonary hypertension of the newborn: population based cohort study from the five Nordic countries. BMJ 344:d8012

Klemetti R, Gissler M, Sevón T, Koivurova S, Ritvanen A, Hemminki E (2005) Children born after assisted fertilization have an increased rate of major congenital anomalies. Fertil Steril 84:1300–1307

Lampi P, Hakulinen T, Luostarinen T, Pukkala E, Teppo L (1992) Cancer incidence following chlorophenol exposure in a community in Southern Finland. Arch Environ Health 47:167–175

Langefors B (1993) Essays on Infology. Summing up and planning for the future. Department of Information Systems, University of Göteborg, Gothenburg

Last JM (1988) A dictionary of epidemiology. International Epidemiological Association, 2nd edn. Oxford University Press, New York/Oxford/Toronto

Li J, Vestergaard M, Obel C, Christensen J, Precht DH, Lu M, Olsen J (2009) A nationwide study on the risk of autism after prenatal stress exposure to maternal bereavement. Pediatrics 123:1102–1107

Lindström P, Janzon L, Sternby NH (1997) Declining autopsy rate in Sweden: a study of causes and consequences in Malmö, Sweden. J Int Med 50:367–375

Ludvigsson JF, Otterblad-Olausson P, Pettersson BU, Ekbom A (2009) The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research. Eur J Epidemiol 24(11):659–667

Ludvigsson JF, Andersson E, Ekbom A, Feychting M, Kim JL, Reuterwall C, Heurgren M, Olausson PO (2011) External review and validation of the Swedish national inpatient register. BMC Public Health 11:450

Manderbacka K, Arffman M, Leyland A, McCallum A, Keskimäki I (2009) Change and persistence in healthcare inequities: access to elective surgery in Finland in 1992–2003. Scand J Public Health 37:131–138

Mattsson B (1984) Cancer registration in Sweden. Studies on completeness and validity of incidence and mortality registers. Thesis. Karolinska Institutet, Stockholm

McDonald CJ, Hui SL (1991) The analysis of humungous databases: problems and promises. Stat Med 10:511–520

Miettunen J, Suvisaari J, Haukka J, Isohanni M (2011) Use of register data for psychiatric epidemiology in the Nordic countries. In: Tsuang MT, Tohen M, Jones PB (eds) Textbook in psychiatric epidemiology, 3rd edn. Wiley-Blackwell, Chichester, pp 117–131

Mortensen LH, Helweg-Larsen K, Andersen AM (2011) Socioeconomic differences in perinatal health and disease. Scand J Public Health 39(Suppl 7):110–114

Motheral B, Brooks J, Clark MA, Crown WH, Davey P, Hutchins D, Martin BC, Stang P (2003) A checklist for retrospective database studies-report of the ISPOR Task Force on Retrospective Databases. Value Health 6:90–97

Nordtveit TI, Melve KK, Albrechtsen S, Skjaerven R (2008) Maternal and paternal contribution to intergenerational recurrence of breech delivery: population based cohort study. BMJ 336: 872–876

OECD, Organisation for Economic Co-operation and Development (2002) Measuring up. Improving health system performance in OECD countries. OECD, Paris

Otterblad Olausson P, Cnattingius S, Haglund B (1999) Teenage pregnancies and risk of late fetal death and infant mortality. Br J Obstet Gynaecol 106:116–121

Paananen R, Gissler M (2011) Cohort profile: the 1987 Finnish Birth Cohort. Int J Epidemiol 41(4):941–945

Parker MJ, Currie CT, Mountain JA, Thorngren K-G (1998) Standardised audit of hip fracture in Europe (SAHFE). Hip Int 8:10–15

Peltola M, Juntunen M, Häkkinen U, Rosenqvist G, Seppälä TT, Sund R (2011) A methodological approach for register-based evaluation of cost and outcomes in health care. Ann Med 43(Suppl 1):S4–S13

Percy C, Muir C (1989) The international comparability of cancer mortality data. Results of an international certificate study. Am J Epidemiol 129:934–946

Pershagen G, Åkerblom G, Axelson O, Clavensjö B, Damber L, Desai G, Enflo A, Lagarde F, Mellander H, Svartengren M, Swedjemark GA (1994) Residential radon exposure and lung cancer in Sweden. New Engl J Med 330:159–164

Potvin L, Champagne F (1986) Utilization of administrative files in health research. Soc Indic Res 18:409–423

PRC (Population Register Centre) (2011) Population information system – Personal identity code. http://www.vrk.fi/default.aspx?id=45. Accessed 8 Jan 2013

Pukkala E, Martinsen JI, Lynge E, Gunnarsdottir HK, Sparén P, Tryggvadottir L, Weiderpass E, Kjaerheim K (2009) Occupation and cancer – follow-up of 15 million people in five Nordic countries. Acta Oncol 48:646–790

Rahu M, Tekkel M, Veidebaum T, Pukkala E, Hakulinen T, Auvinen A, Rytömaa T, Inskip PD, Boice JD Jr (1997) The Estonian study of Chernobyl cleanup workers: II. Incidence of cancer and mortality. Radiat Res 147:653–657

Ringbäck Weitoft G, Gullberg A, Rosén M (1998) Avoidable mortality among psychiatric patients. Soc Psychiatry Psychiatr Epidemiol 33:430–437

Ringbäck Weitoft G, Haglund B, Rosén M (2000) Mortality among lone mothers in Sweden: a population study. Lancet 355:1215–1219

Ringbäck Weitoft G, Hjern A, Haglund B, Rosén M (2003) Mortality, severe morbidity and injury in children living with single parents in Sweden: a population-based study. Lancet 361:289–295

Ringbäck Weitoft G, Ericsson Ö, Jöfroth E, Rosén M (2008) Equal access to treatment? Population-based follow-up of drugs dispensed to patients after myocardial infarction in Sweden. Eur J Clin Pharmacol 64:417–424

Rosén M, Alfredsson L, Hammar N, Kahan T, Spetz CL, Ysberg AS (2000) Attack rate, mortality and case fatality for acute myocardial infarction in Sweden 1987–1995. Results from the Swedish Myocardial Infarction Register. J Internal Med 248:159–164

Sarkola T, Gissler M, Kahila H, Autti-Rämö I, Halmesmäki E (2011) Early health care utilization and welfare interventions among children of mothers with alcohol and substance abuse: a retrospective cohort study. Acta Paediatr 100:1379–1385

Shahar Y (1997) A framework for knowledge based temporal abstraction. Artif Intell 90:79–133

Smith U, Gale EA (2009) Does diabetes therapy influence the risk of cancer? Diabetologia 52:1699–1708

Sørensen HT, Sabroe S, Olsen J (1996) A framework for evaluation of secondary data sources for epidemiological research. Int J Epidemiol 25:435–442

Stenbeck M, Rosén M (eds) (1995) Cancer survival in Sweden in 1961–1991. Acta Oncologica Suppl 4:1–124

Storm HH, Engholm G, Hakulinen T, Tryggvadóttir L, Klint Å, Gislum M, Kejs AMT, Bray F (2010) Survival of patients diagnosed with cancer in the Nordic countries up to 1999–2003 followed to the end of 2006. A critical overview of the results. Acta Oncol 49:532–544

Sund R (2003) Utilisation of administrative registers using scientific knowledge discovery. Intell Data Anal 7:501–519

Sund R (2007) Utilization of routinely collected administrative data in monitoring of aging dependent hip fracture incidence. Epidemiol Perspect Innov 4:2

Sund R (2008) Methodological perspectives for register-based health system performance assessment. Developing a hip fracture monitoring system in Finland. Stakes Research Report 174. National Research and Development Centre for Welfare and Health, Helsinki. http://urn.fi/URN:ISBN:978-951-33-2132-1. Accessed 8 Jan 2013

Sund R, Koski S (2009) FinDM II – On the register-based measurement of the prevalence and incidence of diabetes and its long-term complications. Finnish Diabetes Association, Tampere. http://www.diabetes.fi/files/1167/DehkoFinDM_Raportti_ENG.pdf. Accessed 8 Jan 2013

Sund R, Nurmi-Lüthje I, Lüthje P, Tanninen S, Narinen A, Keskimäki I (2007) Comparing properties of audit data and routinely collected register data in case of performance assessment of hip fracture treatment in Finland. Methods Inf Med 46:558–566

Sund R, Juntunen M, Lüthje P, Huusko T, Häkkinen U (2011) Monitoring the performance of hip fracture treatment in Finland. Ann Med 43(Suppl 1):S39–S46

Sund R (2012) Quality of the Finnish Hospital Discharge Register: a systematic review. Scand J Public Health 40(6):505–515.

Sundgren B (1996) Making statistical data more available. Int Stat Rev 64:23–28

Swedish Local Authorities and Regions (2011) Quality registers. http://www.kvalitetsregister.se/om_kvalitetsregister/quality_registries?UsePrintableVersion=true. Accessed 8 Jan 2013

Tarkiainen L, Martikainen P, Laaksonen M, Valkonen T (2012) Trends in life expectancy by income from 1988 to 2007: decomposition by age and cause of death. J Epidemiol Community Health 66(7):573–578

Teppo L, Pukkala E, Lehtonen M (1994) Data quality and quality control of a population-based cancer registry. Acta Oncol 33:365–369

Tibblin G, Wilhelmsen L, Werkö L (1975) Risk factors for myocardial infarction and death due to ischaemic disease and other causes. Am J Cardiol 35:514–522

Tiihonen J, Lönnqvist J, Wahlbeck K, Klaukka T, Niskanen L, Tanskanen A, Haukka J (2009) 11-year follow-up of mortality in patients with schizophrenia: a population-based cohort study (FIN11 study). Lancet 374:620–627

Tretli S, Engeland A, Haldorsen T, Hakulinen T, Hörte LG, Luostarinen T, Schou G, SigvaldasonH, StormHH, Tulinius H,Vaittinen P (1996) Prostate cancer – Look to Denmark? J Natl Cancer Inst 88:128

Vågerö D, Persson G (1987) Cancer survival and social class in Sweden. J Epidemiol Community Health 41:204–209

Wahlbeck K, Westman J, Nordentoft M, Gissler M, Munk Laursen T (2011) Outcomes of Nordic mental health systems: life expectancy of patients with mental disorders. Br J Psychiatr 199:453–458.

Walsh PC (2002) Overdiagnosis due to prostate antigen screening: lessons from U.S. prostate cancer incidence trends. J Natl Cancer Inst 94:981–990

Emergency and Disaster Health Surveillance

21

Susan T. Cookson and James W. Buehler

Contents

S.T. Cookson
CAPT, US Public Health Service, Center for Global Health, Centers for Disease Control and Prevention, Atlanta, GA, USA

J.W. Buehler (✉)
Public Health Surveillance Program Office, Office of Surveillance, Epidemiology, and Laboratory Services, Centers for Disease Control and Prevention, Atlanta, GA, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_61,

21.1 Introduction

During crises that threaten public health, as during more ordinary circumstances, surveillance is a fundamental part of public health practice and is used to measure the impact of disease, detect changes in trends, guide immediate and long-term actions, and prioritize the use of public health resources. Surveillance is especially critical during times of crises such as epidemics, natural disasters, famines, and conflicts. During events that rise to the level of becoming humanitarian emergencies, the highest rates of mortality and morbidity occur during the early phases when resources are scarce, healthcare and other service delivery systems are disrupted, and the health, safety, security, and well-being of populations are at risk (Toole and Waldman 1990). Compared with routine surveillance, surveillance during major emergencies is abbreviated in order to focus on those diseases or conditions with epidemic potential or with the highest risk of severe morbidity or mortality. During other urgent but less severe public health events, such as outbreaks or disasters that have lesser impacts, surveillance is needed to characterize health threats, track the health of affected populations throughout the course of events, and complement more targeted epidemiological and laboratory investigations. Absent clearly imminent threats to public health, certain surveillance methods can be employed on an ongoing basis to provide early indication or rapid characterization of disease outbreaks, should they occur, enabling prompt investigations and responses. Across this spectrum of scenarios, a common theme is the importance of the timeliness of surveillance, including the need to collect, manage, analyze, interpret, and report surveillance information quickly in order to be useful in urgent and rapidly evolving situations. This premium on timeliness may constrain the ability of epidemiologists to maximize other desirable attributes of surveillance, such as detailed information, completeness, or specificity.

As Dr. William Foege stated over 35 years ago, “The reason for collecting, analyzing and disseminating information on a disease is to control that disease. Collection and analysis should not be allowed to consume resources if action does not follow. Appropriate action, therefore, becomes the ultimate response goal and the final assessment of the earlier steps of a surveillance system” (Foege et al. 1976). Besides detecting and responding to epidemics, additional objectives of public health in disasters and emergencies, as in routine times, are to determine the healthcare priorities requiring intervention, target the available resources, evaluate the effectiveness of programs that are existing or newly established to determine if they are meeting those priorities, and reassess the priorities over time.

This chapter provides an overview of the current field of disease surveillance during emergencies, disasters, and other urgent population health crises and of surveillance approaches used to detect or rapidly characterize outbreaks. After defining terms, the chapter examines different mechanisms for collecting surveillance data and for communicating that information during urgent situations. The chapter will contrast surveillance methods used in developed and developing countries and give current examples where they have been deployed. Disaster planning and emergency preparedness efforts should take into account the full spectrum of potential threats that communities might face and the attendant surveillance information needs of public health emergency and humanitarian relief responders. For a more general overview of the principles of public health surveillance, (see chapter ▶Infectious Disease Epidemiology of this handbook).

21.2 Definitions of Events

21.2.1 Disease Outbreaks

Outbreaks represent an increase beyond expected in cases of disease within a given geographical area or group of people over a particular period of time. The terms "outbreak" and "epidemic" are often used interchangeably, although "epidemic" is often reserved to describe events that affect relatively large numbers of people (Centers for Disease Control and Prevention 2004). The term "epidemic" may be used to describe public health problems that evolve over periods of years, e.g., the "obesity epidemic," but such use is beyond the scope of this chapter. Outbreaks are rarely triggered intentionally, but increasing concerns about the threat of bioterrorism since the early 2000s have heightened investments in surveillance systems aimed at the early detection and prompt characterization of outbreaks arising from intentional acts. The earlier disease outbreaks are recognized, the more quickly investigations can be conducted and control measures initiated, with the attendant benefits of minimizing morbidity or mortality. Outbreaks may arise from exposures to infectious agents, toxins, or other environmental hazards. Humanitarian emergencies, natural disasters, and mass gatherings might give rise to circumstances associated with increased risks for disease outbreaks, although vulnerabilities and potential impacts are highly variable across different situations.

21.2.2 Mass Gatherings

High-profile events, such as sporting championships, political conventions or demonstrations, religious gatherings, or other events that attract large numbers of people, may stress public health infrastructures in ways that increase the vulnerability for a disease outbreak. These events might also be targets for a bioterrorist attack. Mass gatherings may necessitate efforts to bolster existing public health surveillance capacity or initiate special surveillance activities for a time-limited period before, during, and after the event. In these situations, surveillance planning needs to take into account the potential for threats to public health, which may or may not arise. Surveillance planning should consider the unique challenges posed by the influx of large numbers of people who are not normally residents of an area, including plans for the delivery of healthcare services during the event and associated impacts on the ability of healthcare facilities to provide surveillance data during periods when staff and other resources might be stressed. During mass gatherings at high-profile forums, multiple governmental agencies and non-governmental organizations (NGOs) are likely to be engaged to address public safety and health needs, including agencies and organizations that would not normally be convened in a particular locality. As a result, procedures for reporting and use of surveillance information should be established as part of larger advance planning activities (Stergarchis and Tsouros 2007). These mass gatherings are usually defined as having more than 25,000 participants (Lombardo et al. 2008) and often include international sporting games.

21.2.3 Natural Disasters and Humanitarian Emergencies

Although often used interchangeably, the terms "natural disaster" and "humanitarian emergency" (or "complex emergency") represent events with different causes and consequences, with the former resulting from ecological events and the latter from the breakdown of governance and authority. The Pan American Health Organization (PAHO) defines natural disasters as overwhelming ecological disruptions of sufficient magnitude that assistance is usually required from outside agencies and organizations (Pan American Health Organization 1981). The United Nations Disaster Reduction Secretariat declared 2010 as the deadliest year in the past two decades based on data compiled by the Centre for Research on the Epidemiology of Disasters. In 2010, there were over 370 natural disasters, 300,000 people killed, and nearly 208 million persons affected, with the top two lethal disasters being the Haiti earthquake and the Russian heat wave (United Nations Radio 2010). Natural disasters can occur with varying degrees of advance warning and can be of atmospheric or geologic origin, including severe weather in the form of floods, droughts, heat waves, ice storms, and hurricanes/cyclones or geological events, such as earthquakes, tsunamis, volcanic eruptions, avalanches, and landslides. They usually occur quickly (droughts being an exception) and have the greatest impact on lives in the first few days following the event.

The Inter-Agency Standing Committee (IASC) was formed in 1992 to strengthen coordination, policy development, and decision-making among United Nations agencies and partner organizations in responding to complex humanitarian emergencies (Inter-Agency Standing Committee 2009). The IASC defines a complex emergency as a "a humanitarian crisis in a country, region or society where there is total or considerable breakdown of authority resulting from internal or external conflict and which requires an international response that goes beyond the mandate or capacity of any single and/or ongoing UN country program" (World Health Organization 2011a). Commonly used mortality thresholds in defining humanitarian emergencies include a crude mortality rate of one death or more per 10,000 total population per day or two or more deaths per 10,000 children under 5 years of age per day (Checchi and Roberts 2005). In areas where the baseline mortality rates are known, the threshold for humanitarian emergencies is a doubling of those rates. These indicators also are useful for monitoring the extent to which relief services are meeting the needs of the affected population and, thus, the overall impact of the relief response.

Natural disasters and humanitarian emergencies typically result in different patterns of morbidity and mortality. In natural disasters, injuries are the leading cause of death, including deaths due to blunt trauma or crushes, drowning, or hyper- or hypothermia, while deaths due to infectious disease are less common (Watson et al. 2007). In contrast, infectious disease epidemics are common during complex emergencies. For example, in the 10-year period of 1995–2004, infectious disease epidemics occurred during 63% of the 30 largest complex emergencies versus 23% of the 30 largest natural disasters (Spiegel et al. 2007). Contributors to epidemics during complex emergencies include the lack of authority, security, healthcare infrastructure, basic services, and access to and for the population because of security issues, as well as vulnerabilities arising from human rights violations; marginalization of specific ethnic, religious, social, or political groups; and mass displacements. Mass displacements following from natural disasters can also lead to outbreaks depending on the size of displaced populations, access to potable water, and sanitation (Spiegel et al. 2007), as seen in the occurrence of outbreaks of gastrointestinal illness within days in Muzaffarabad following the 2005 Pakistan earthquake (United Nations Country Team in Pakistan 2005). Because they were homeless from the earthquake, about 10,000 people set up spontaneous camps with an unsafe water supply and poor sanitation facilities. Without separate latrines for those with acute watery diarrhea and with a water system damaged by the earthquake, further contamination of the water supply resulted. The international humanitarian community responded by supporting the Pakistan Government with water supply repair, adequate sanitation facility construction, and provision of emergency health care to those with diarrhea and dehydration.

21.3 Formal Surveillance

Surveillance in the context of disasters or other urgent situations may depend on enhancements to routine surveillance systems not specifically designed to serve emergency preparedness and response activities, ongoing surveillance systems specifically designed to provide very timely information for outbreak detection or characterization, event-prompted activation of a previously developed emergency surveillance or assessment utility, or ad hoc implementation of an event-specific surveillance system. In crises, as in more routine surveillance practice, surveillance may draw on information generated when people use healthcare services, such as hospital emergency departments or emergency field clinics, or on population surveys conducted to assess health status. A spectrum of surveillance methods may be employed, geared to the information needs of a specific situation and the resources available to conduct surveillance.

21.3.1 Uses of Surveillance

Among questions that surveillance systems might be tasked to address, three essential questions regarding disease outbreaks include the following:

Is an outbreak occurring? Outbreaks are detected in a variety of ways, depending on a mix of disease-specific, epidemiological, population, and other factors (Buehler et al. 2003). If sufficient numbers of cases of disease are manifest within a community, the outbreak will be apparent to clinicians as an unusual increase in patients with a particular pattern of disease. Many outbreaks are detected in this way, including outbreaks associated with previously unknown diseases as well as with familiar causes of food- or waterborne disease. However, in emergencies or disasters with mass migration, an increase in case numbers might be caused by an increase in the population size due to population movement. In these situations, knowing migration patterns and receiving or obtaining periodic counts of the affected population are critical to understanding changes in disease surveillance case numbers and determining if true outbreaks are occurring. In other instances, the emergence of disease might be too geographically diffuse or unfold too slowly to be recognizable by individual clinicians. In these instances, different types of disease surveillance systems might provide the initial indication of an outbreak. This surveillance includes traditional systems based on mandated reporting of diagnosed cases of "notifiable" diseases; automated systems that monitor trends in the manifestations of disease – "syndromic surveillance" systems that monitor respiratory, gastrointestinal, and other syndrome-based categories of disease (Buehler et al. 2003; Mandl et al. 2004); or laboratory-based systems that track particular attributes of microbiologic agents, such as bacteriological serotypes or DNA "fingerprints" (Swaminathan et al. 2001). Selected notifiable diseases are aggregated weekly from the 50 US states and the 5 US territories and reported by the Centers for Disease Control and Prevention (US CDC). These include both infrequently reported diseases such as anthrax, measles, and congenital syphilis and more frequent diseases such as chlamydia, Lyme disease, and varicella. Regular examination of surveillance data, including examination of alerts generated by statistical algorithms used to detect aberrant temporal or temporal/geographical trends, might signify the emergence of an outbreak. The most common use of syndromic surveillance data by US health departments is to detect the beginning and end of the annual influenza season (Buehler et al. 2008).

In yet other instances, the diagnosis of a single case of a rare disease might be an indication of a potential disease outbreak, such as the original recognition of inhalation anthrax infection in an employee of a media company in 2001 in the United States (Traeger et al. 2002) or a single case of measles, regardless of location. In settings where clinical laboratory testing is not routinely available, targeted laboratory testing of clinical specimens can support the recognition of epidemics, as in the detection of shigellosis in Sudan in 2004 (Musani et al. 2004). Any such indication should prompt an examination of data from existing surveillance systems to confirm or exclude the presence of an outbreak. Depending on the situation, the review of surveillance data combined with outreach to local healthcare providers can improve their awareness of possible unusual trends in illness. Conversely, surveillance can be useful to determine that an outbreak is not occurring, either to provide assurance that suspect case reports are not harbingers of a more widespread problem or to quell rumors during a crisis.

What are the characteristics of recognized outbreaks? Regardless of how outbreaks are detected, once they are recognized, there is a need for timely monitoring to track their course from beginning to end, often requiring enhancements to routine surveillance procedures or special ad hoc surveillance efforts. Surveillance may be coupled with targeted epidemiological and laboratory investigations to characterize the extent of exposures to infectious or toxic agents, the extent and severity of disease and accompanying needs for healthcare services, the impact and effectiveness of interventions, the need for immunization (e.g., determine if available vaccine protects against serotype of meningococcal outbreak), or the potential adverse effects of medications or vaccines provided to mitigate the impact of an outbreak. The need for information often evolves over the course of an event. Intensive efforts to identify and characterize individual cases of disease might be appropriate at the initial stage of an event, but such efforts might not be feasible or necessary at later stages of an event. If very large numbers of cases are occurring, information priorities are likely to shift to tracking the broader sweep of an epidemic, requiring less detailed or specific information. For example, following initial recognition of illness due to H1N1 influenza A in the United States in 2009, initial surveillance efforts focused on reporting individual cases and were accompanied by investigations to characterize the clinical, laboratory, and epidemiological features of the disease (Swerdlow et al. 2011). As familiarity with the new pandemic strain of influenza increased and as infections became widespread, less specific measures were employed to monitor the sweep of the pandemic across the country, drawing on procedures developed over many prior influenza seasons to monitor "influenza-like illness" (Brammer et al. 2011).

Are anticipated health complications occurring following disasters? Experience with prior disasters can be used to anticipate possible health consequences, such as injuries associated with exposure to physical forces or cleanup efforts, carbon monoxide exposures associated with inappropriate venting of gasoline-powered generators in situations where electricity service is disrupted, or infectious diseases in situations where there is mass displacement, where sanitation services are strained, or where water and supplies are contaminated or compromised. Disruption in access to routine healthcare services can lead to an increase in complications of chronic diseases, such as diabetes or cardiovascular diseases, or infectious diseases requiring long-term treatment, such as tuberculosis or HIV infection (Mokdad et al. 2005). Large-scale disasters with substantial morbidity and mortality or with environmental impacts that affect the economic or social well-being of communities can have substantial impacts on mental health, which may manifest both as symptoms of psychiatric illness and as somatic disease (Gerrity and Flynn 1997; Green et al. 2006; Thienkrua et al. 2006; van Griensven et al. 2006). Surveillance systems aimed at tracking various categories of illness may be complemented by rapid community needs assessment surveys that aim to provide a more comprehensive assessment of the health needs of populations affected by disasters, such as the Community Assessment for Public Health Emergency Response (CASPER) toolkit developed by the US CDC (Centers for Disease

Table 21.1 Sectors evaluated in the Initial Rapid Assessment

Sector names	Number of subsections within sectors
Population	7 plus any additional information (AI)
Shelter and essential non-food items	5 plus AI
Water supply, sanitation, and hygiene	6 plus AI
Food security and nutrition	10 plus AI
Health risks and status	4 plus AI
Health facility/outreach site[a]	3 plus AI

[a]One per facility or site visited

Control and Prevention 2009) for developed countries and the IASC Initial Rapid Assessment tool for global use (Inter-Agency Standing Committee 2009). The Initial Rapid Assessment (IRA) looks at six sectors (Table 21.1) that need evaluation following a humanitarian emergency or large-scale disaster. The four subsections for the health facility/outreach site sector that are assessed include (1) general information of the facility such as location, contact person, type of facility (hospital, center, post, outreach), level of damage and patient access, and financial requirement of patients; (2) resources such as staffing types and numbers and essential drugs, vaccine, and simple supply availability; (3) any additional information and (4) type of care provided such as community outreach; primary care including feeding program, HIV, mental health; and secondary care with surgery, laboratory, X-ray, and blood bank services.

21.3.2 Enhanced Notifiable Disease Surveillance

In some instances, there may be a premium on ensuring that all cases of a highly critical disease are identified during an outbreak, such as a situation when there is a need to contact every affected person to interrupt transmission of a transmissible infection associated with severe morbidity. One approach would be to enhance routine procedures for notifiable disease surveillance (i.e., surveillance for conditions where healthcare providers are mandated to report cases to public health authorities). This occurred in 2001 in the United States when cases of inhalational and cutaneous anthrax occurred in several communities as a result of intentional contamination of mail with anthrax spores. For example, in the Trenton, New Jersey, region, where contaminated letters entered the postal system and where anthrax infections were detected, an extremely active approach to surveillance for possible anthrax cases was implemented. This involved daily outreach to area hospitals to identify patients who met a relatively broad definition for possible inhalational anthrax, followed by case reviews, and, when indicated, laboratory testing for anthrax infection. This labor-intensive approach to surveillance led to assurances that the outbreak had ended and was shortly discontinued since such intensive efforts on the part of hospital and public health personnel were no longer necessary or sustainable (Tan et al. 2002).

21.3.3 Syndromic Surveillance

Surveillance of disease syndromes – constellations of symptoms or other manifestations of disease – is a long-standing element of public health surveillance, especially when the etiology of a newly recognized disease is unknown and surveillance definitions must depend on syndromic criteria, when resources are not available to routinely diagnose disease or the ability of the public health laboratory system is inadequate, or when definitive diagnostic testing is not routinely necessary or used in practice for common illnesses. Nonetheless, the term "syndromic surveillance" has been applied to relatively new surveillance systems specifically geared to rapid monitoring on a daily or more frequent basis of syndromic categories of disease, prompted in large part by concerns about the threat of bioterrorism since 2001. While investments in "syndromic surveillance" in developed countries has incorporated the use of electronic health records and automation of key steps in the surveillance process, syndromic surveillance can also involve surveillance that depends on manual data collection. When automated healthcare information systems are available and accessible, the application of informatics tools to tap and extract selected routinely recorded data from these systems obviates the need for clinicians or healthcare staff to modify their activities for surveillance purposes. Automated information management tools can be further applied to collect such information from multiple facilities; compile the data into a common data store; scan recorded descriptions of patients' symptoms or diagnoses and collate these into counts of patients with various categories of disease, such as respiratory illness, upper or lower gastrointestinal illness, or influenza-like illness; perform statistical tests for aberrant temporal or temporal/geographical trends; and automatically display the results for review by epidemiologists. Such systems can be used to scan for possible disease outbreaks, characterize community illness patterns when outbreaks are suspected for any reason, track the course of relatively large outbreaks, and monitor expected seasonal illness due to influenza, other respiratory viruses, and common viral gastrointestinal infections, such as norovirus or rotavirus (Heffernan et al. 2004; Mandl et al. 2004). In the event of a disaster, syndrome categories might be modified or attention honed to track potential infectious and non-infectious complications or injuries.

A common approach to syndromic surveillance is to monitor patient visits to hospital emergency departments, although a large mix of sources have been applied, including outpatient clinics, calls to ambulance dispatch systems, calls to services that provide telephone guidance to patients, and calls to poison control centers (Heffernan et al. 2004; Mandl et al. 2004; Buehler et al. 2008). One such example was the monitoring of patient emergency department visits in Georgia by residents from Louisiana and Mississippi following Hurricane Katrina (Cookson et al. 2008). Of residents who visited participating emergency departments, 174 visited during the 8 months before and 151 visited during the 1 month after Hurricane Katrina, reflecting a sevenfold increase in monthly visits. Cardiopulmonary complaints significantly increased ($p = 0.03$) after Hurricane Katrina. Examining emergency department data as they are received can help assess needs among disaster victims.

Syndrome categories may be based on algorithms that scan the recorded text of patients' self-description of their symptoms, i.e., the "chief complaint"; recorded diagnoses, including diagnoses coded using the International Classification of Diseases (ICD); or both. Other approaches depend on the use of services, such as the purchase of over-the-counter medications, or other manifestations of illness, such as school or workplace absenteeism.

A wide variety of statistical techniques have been developed to detect aberrant trends – an increase beyond expected in the number of patients falling into a particular syndrome category, with varying levels of complexity being incorporated to account for underlying patterns in the use of healthcare services, such as variations in the use of care by day of the week, and to consider geographical as well as temporal clustering of disease (Buckeridge et al. 2005). In practice, relatively simple approaches not requiring extensive historical data are commonly used, such as the Early Aberration Reporting System, based on the cumulative sum or CUSUM statistical method (Centers for Disease Control and Prevention 2010a). CUSUM uses a running baseline or denominator that consists of the average number of counts for the selected syndrome from a previous 7-day period. This denominator is compared with the numerator of current syndrome counts. Traditionally, three different CUSUM methods are performed using different numerator and denominator counts. If the result of one of these three CUSUM methods is significantly greater than the expected rate, a signal is generated.

When syndromic surveillance is conducted for a short period of time, as might be necessary during an urgent response to a crisis or for a time-limited period shortly before, during, and after a high-profile public event or mass gathering, the term "drop-in" surveillance might be used to describe the work of a team mobilized to implement and support a surveillance system for the duration of the event. "Drop-in" surveillance may depend on hospital or emergency department staff or surveillance team members to record the necessary surveillance data manually if automated methods cannot be applied, although dependence on such manual effort is not likely to be sustainable beyond the period of the urgent situation (Centers for Disease Control and Prevention 2002a).

21.4 Communication

Surveillance systems are often based on the roles and relationships among healthcare providers and public health officials within communities and the ability to use lines of communication arising from these relationships is essential during crises. Formal surveillance systems may be complemented by enhanced outreach by public health staff to key contacts at clinical facilities or by specially convened in-person, teleconference, or web-enabled meetings during crises. These engagements can foster the exchanges that enhance insights into the status of an outbreak or a population's health that can be more nuanced than insights available from surveillance data alone. During emergencies, such communication among response participants, either formal or informal, is critical, and various formats for communication may be used in different situations.

21.4.1 Formal Systems: Inter-Agency Standing Committee, Cluster Approach

After the December 2004 Indonesian tsunami, humanitarian organizations and agencies cited the need to improve the leadership and coordination of aid responses in large disasters or complex emergencies. In response, the IASC developed the "cluster approach" to clarify roles, responsibilities, and lines of authority and delegation, with the aim of strengthening partnerships among host ministry of health NGOs, international organizations, the International Federation of the Red Cross/Red Crescent (IFRC), and United Nations agencies (World Health Organization 2006). The cluster approach was established to address gaps found in the effectiveness of humanitarian response through building partnerships. Its purpose is to ensure predictability and accountability of the international responses to humanitarian emergencies by clarifying the division of labor among organizations, defining the roles and responsibilities of the different sectors of the response, and making the international humanitarian response more structured and accountable to the need of the partner host governments, local authorities, and local civil society. At the global level, the objective of the cluster approach is to strengthen system-wide preparedness and technical capacity, while at the country level it is to improve predictability of leadership and accountability among partners across all "sectors" of response activities. During a large-scale natural disaster or complex emergency, for countries with designated humanitarian coordinators (Table 21.2), the cluster approach is to be activated. These countries have been appointed a humanitarian coordinator because they can and do face sudden major emergencies requiring a multisectoral response with the participation of a wide range of international humanitarian responders. At the daily to weekly health cluster meetings occurring once the emergency has arisen and the cluster has been activated, information is shared by the partner host ministry of health, different participating NGOs, international organizations, IFRC, and United Nations agencies. At the meetings, such information as standardized case definitions and surveillance forms, methods and routes for specimen collections, and increases in cases of outbreak potential syndromes or diseases are shared.

Among the cluster approach response sectors, three – water, sanitation, and hygiene (WASH); nutrition; and health – involve public health directly, while other sectors address capacities or services that can affect health, such as emergency shelter and protection for displaced people and telecommunications. Websites, such as ReliefWeb (2013) (http://www.reliefweb.int) and Humanitarian Response (2013) (http://www.humanitarianresponse.info) (formerly OneResponse), have been created to be a source of information and to improve coordination during humanitarian responses. Preliminary assessments of responses to the 2010 earthquake and subsequent cholera epidemic in Haiti illustrated the opportunities available in communication by implementing the cluster approach (Pan American Health Organization 2010). In Haiti, health and WASH cluster meetings provided a venue for in-person sharing of information among representatives of organizations participating in the response, including NGOs providing healthcare services in camps for displaced people, and for coordinating surveillance efforts, including the establishment of a

Table 21.2 Countries by United Nations regions with designated humanitarian coordinators using the cluster approach, 2011[a]

Middle East and North Africa	**Africa**
	East and Horn of Africa
Iraq	Ethiopia
Occupied Palestinian Territory	Kenya Somalia
Yemen	Sudan
Asia and Pacific	Uganda
Afghanistan	**Central and Great Lakes**
Indonesia	
Nepal	Burundi
Pakistan Sri Lanka	Central African Republic
Timor-Leste	Chad
The Americas Colombia	Democratic Republic of Congo
Haiti	**West**
	Côte d'Ivoire
	Guinea Liberia
	Niger
	Southern
	Zimbabwe

[a] Adapted from the Health Action in Crisis, Cluster Countries (World Health Organization 2011b)

surveillance system for monitoring health in the camps and subsequently tracking cholera cases. In addition, the Haitian Ministry of Public Health and Population in collaboration with PAHO and US CDC created an Internet-based forum to share information, such as disease case definitions for surveillance, surveillance forms, and information on observed disease trends, including cholera (Centers for Disease Control and Prevention 2010b, c). Because of high turnover of NGO staff and because partners cannot always attend cluster meetings as clinics are located across a wide geographical area and to obtain definitions and forms, these types of publicly accessible Internet forum (ReliefWeb, Humanitarian Response, or Haiti Internet-based forum) allow for the exchange of messages and materials either with the entire group or a specific staff member. Electronic files can be posted or downloaded by any member. The Haiti Internet-based forum was effective in encouraging timely and reciprocal communication between surveillance coordinators and NGOs; weekly feedback reports were available to NGOs that included analysis of trends of proportions of each reportable condition. This probably led to the high response rate from the large clinics seen across the weeks; from epidemiological week 5–16, a mean and median of 35 clinics were reporting weekly (range: 12–48).

21.4.2 Incident Command Structures

Increasingly, public health agencies in the United States are implementing the use of formal systems of "incident command" or "incident management" to manage public health activities during crises and to facilitate communications with public health colleagues in neighboring jurisdictions or different levels of government (local, state, or national), healthcare providers, and other responder agencies and organizations. Systems of incident command have a longer history of use by public safety and emergency management agencies. By training and acquiring skills in the use of incident management protocols, public health agencies are better prepared to interact with these agencies, which typically have lead responsibility for managing multiagency emergency response activities. For example, the National Incident Management System provides standardized approaches for organizing, leading, and managing emergency response activities and for channeling information, such as data from surveillance systems (Federal Emergency Management Agency 2007). One such structure for leadership and management used by the US CDC for their support to the Haitian government following the 2010 earthquake is illustrated below (Fig. 21.1). Pre-event, multiagency participation in exercises that use this approach can minimize confusion when participants from different agencies and disciplines, each with different cultures and jargon, collaborate during crises. For example, the term "surveillance" itself describes very different activities when used by epidemiologists and law enforcement investigators who may be collaborating in responding to a possible bioterrorist threat.

21.5 Additional Tools for Public Health Response

In responding to large-scale disasters or humanitarian emergencies, multiple surveillance approaches and other tools that complement public health surveillance might be used at different stages of the response, including initial rapid assessments, facility-based health surveillance, and surveys for specific diseases or conditions (Table 21.3). To quickly assess basic needs for shelter, water, food, sanitation, and health care immediately after an event, the first tool used is often a rapid assessment of the extent of destruction or disruption of essential services. Ad hoc response without a formal needs assessment can lead to inappropriate and ineffective humanitarian responses. To support rapid assessments, the IASC and global leads for the WASH, health, and nutrition clusters have developed and continue to refine the IASC Initial Rapid Assessment tool (Inter-Agency Standing Committee 2009). In 2010, this was used following the earthquake in Haiti and during the monsoon floods in Pakistan. For the Pakistani floods, its purpose was to provide a rapid overview of the emergency situation, identify the immediate impacts of the floods, estimate needs of the affected population for assistance, and define the humanitarian response priorities (and needed funding) in the early weeks of the crisis (World Health Organization 2010a). A total of 515 councils in eight

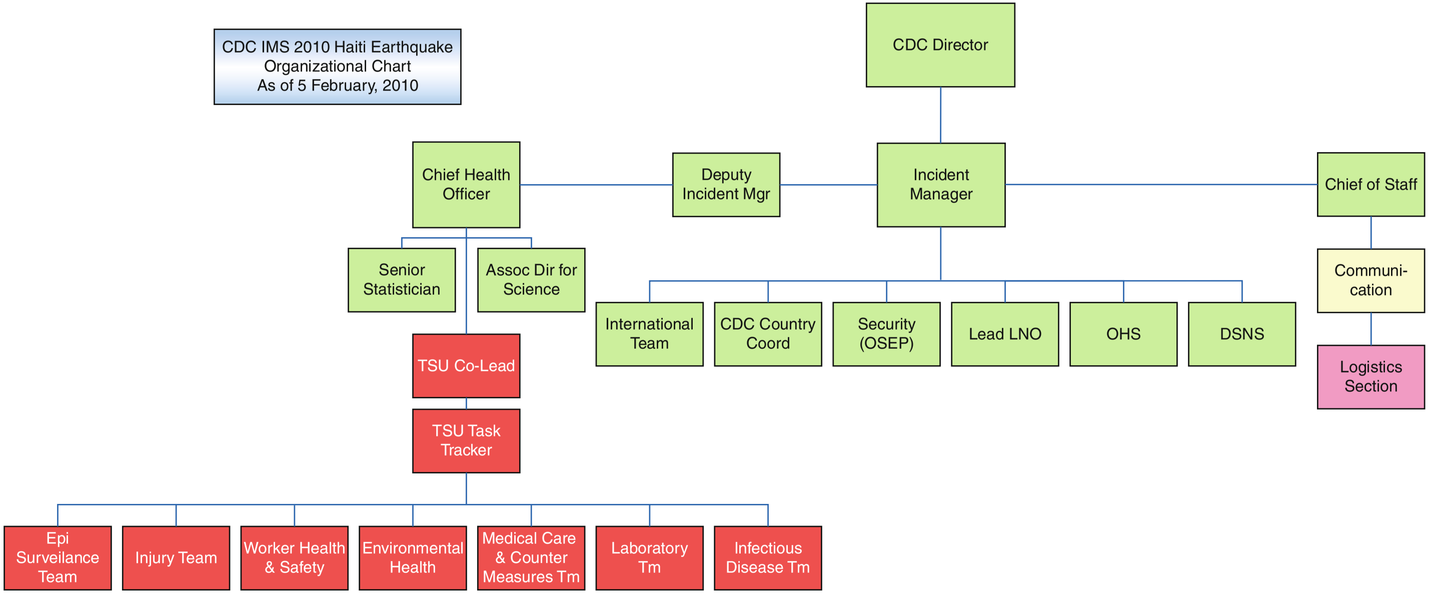


Fig. 21.1 Centers of Disease Control and Prevention (CDC), Incidence Management Structure (IMS) for support and response to 2010 Haiti earthquake organizational chart, 5 February 2010. Legend: *TSU* Technical Support Unit, provides scientific expertise for response in areas of epidemiology, laboratory, and other disease-specific expertise, *OSEP* Office of Security and Emergency Preparedness, coordinates security with other local law enforcement for CDC personnel at an event site and coordinates personal (security clearance) and the physical security of CDC campus, *LNO* Liaison Officer, coordinates with other agencies in providing emergency assistance, *OHS* Office of Health and Safety, ensures that personnel have preventive health measures and assesses resiliency of those deployed, *DSNS* Division of Strategic National Stockpile, provides medical counter measures, if needed

Table 21.3 Types of tools, their characteristics, and methods for public health response to large disaster or complex emergency

	Tool		
Characteristics	Rapid assessment	Survey	Surveillance
Purpose	Preliminary overview	Cross-sectional in-depth appraisal	Detect outbreaks and monitor trends in diseases or conditions of public health importance
Data quantity	Wide variety	Wide variety if sound methods are not followed	May be limited due to constraints and service demands affecting healthcare providers
Source	Observations, often convenience based	Sample of population	Case reports from all or subset of care providers (selected to be as representative as possible)
Time frame	Rapid, single time	Medium term, single time after other tools are completed or established	Ongoing over course of event
Data method and type	Qualitative/review of secondary sources	Quantitative/cross-sectional view/ numerator and denominator based on sample design	Quantitative/numerator and (if possible) denominator (usually proportional rate, using total number of cases)

flood-affected districts in Punjab were assessed. The assessment teams were without any specialized technical expertise and were of both gender, with female assessors conducting group discussions with women and children. Health staff was assigned the role of monitoring and supervising of the teams during data collection by providing on-the-spot assistance to each assessment team. The data were collected and reported in 24–48 h per council, and data for all eight districts were completed in eight days. Rapid assessments should be performed as quickly as possible to identify the affected region and population, set immediate response priorities, determine resource needs and identify essential health information needs. Rapid assessment findings should enable governments, the United Nations agencies, and NGOs to gauge their in-country capacities relative to observed needs, as well as inform aid donors. These rapid needs assessments should combine available information on services and health status to define the pre-event context of the disaster with observations of the post-event situation. Interviews with key informants, such as village or religious leaders; informal focus group discussions with affected persons, and visits to a convenience sample of households are additional methods that can be used. While theoretically desirable, efforts to collect representative data based on statistically valid sampling methods might not be practical or implemented quickly and can waste valuable time (Garfield et al. 2011).

A major challenge in rapid assessments is defining the size of affected persons since individuals and families may be constantly moving to escape violence or to fill needs. One of the best conducted rapid assessments was during the Pakistan 2010 floods. In spite of that, capturing population data are still difficult. Of the eight assessment districts, only 27% had a registration process in their communities to track the affected populations. In Haiti, the International Organization for Migration developed a Displacement Tracking Matrix (DTM) to create a comprehensive list of all known and assessed settlements (Camp Coordination Camp Management Cluster in Haiti 2011). As these types of registrations become more used, the quality of affected-population data will continue to improve. These population figures are needed both to assess whether mortality rates reach emergency thresholds and to determine needs for food, water, shelter, and non-food items, such as cooking items, water storage containers, and soap. Frequent constraints in performing rapid assessments include difficulties in reaching affected areas because of security concerns, problems with logistical support, and transportation barriers. Because of potential biases in available data and differences in perspective of key informants, including the potential for the intentional introduction of misinformation, it is important to consider and synthesize information from as many sources and perspectives as possible. To be useful, the findings of the rapid assessment must be reported promptly to be relevant, given the potential for postdisaster events to unfold rapidly; the recommendations should be concise and specific to the roles of responsible partners and agencies; and the limitations should be clear.

After initial rapid assessments, population health surveys are a tool that is often used during emergencies. For health status measures and conditions that are not readily amenable to accurate assessment through healthcare facility-based surveillance, such as mortality, malnutrition, and gender-based violence, surveys using a representative sample design might be the method of choice to obtain population-based estimates. Systematic random sampling is the preferred method for conducting surveys; however, organized lists of the population or well-organized layout of the camps are not often the reality. Therefore, multistage cluster design sampling is the method often used for conducting these surveys, although this relatively complex method is susceptible to measurement errors if not done correctly. An assessment of 67 surveys performed over 15 months in relation to the famine in Ethiopia in 1999–2000 found that major methodical errors occurred with the vast majority of the surveys, resulting in the possibility that unreliable findings might have influenced policy and resource allocation decisions (Spiegel et al. 2004). Of the 125 surveys evaluated, only 52% used the accepted at the time standard cluster design of 30 clusters × 30 households per cluster. Of these 52%, over 80% used non-random sampling, and almost 10% used sample sizes less than 500 children. Because of the potential for such errors, in 2002, an interagency initiative, the Standardized Monitoring and Assessment of Relief and Transitions (SMART) project, was begun to improve survey methods for three key indicators of the severity of humanitarian crises: the nutritional status of children under 5 years of age and mortality rates for the total affected population and children under 5 years of age. A survey manual and accompanying analytical software were developed.

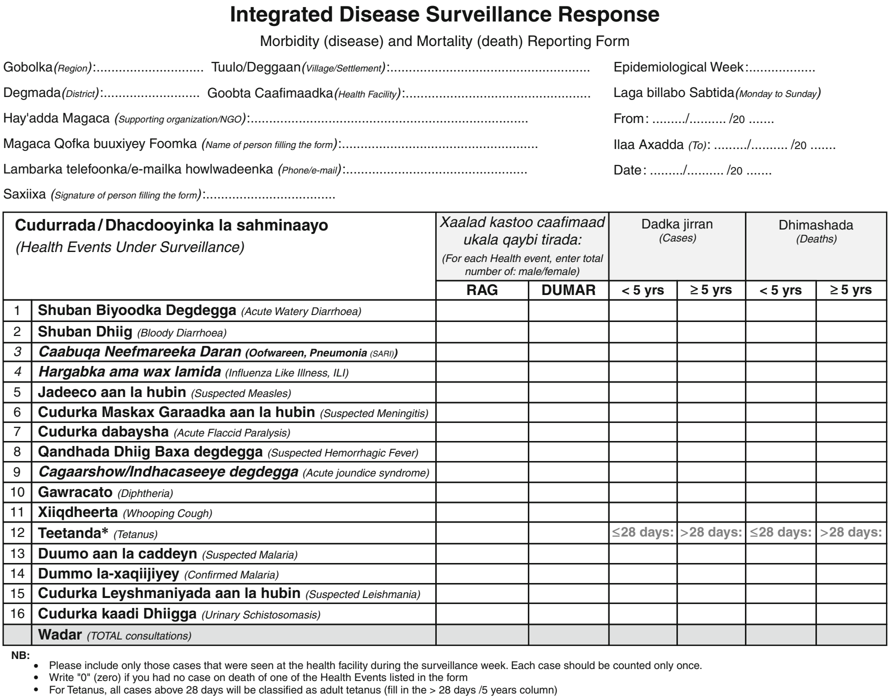
The survey manual was designed to be used by field staff, persons who may have limited understanding and knowledge of epidemiology and statistics. SMART was initiated to improve the technical capacity of agencies to systematically conduct, analyze, interpret, and report survey findings in a standardized and reliable manner. Training sessions have been conducted throughout the world with field staff before and when crises are unfolding such as the 2011 famine in the Horn of Africa. The goal of SMART is to "make the survey process as easy as possible for the field staff and as reliable as possible for the decision-makers" (Standardized Monitoring and Assessment of Relief and Transitions 2009). The software used for SMART can be found on the SMART website (http://www.smartmethodology.org).

21.6 Surveillance in Less-Developed Countries

Surveillance systems are established to detect outbreaks and monitor trends in specific diseases or conditions of public health importance (Médecins Sans Frontières 1997), even though the risk of epidemics is lower in disasters compared with complex emergencies (Spiegel et al. 2007). Rumors of outbreaks and unconfirmed reports occur frequently following disasters and during humanitarian emergencies or mass gatherings. A well-functioning surveillance system can reassure the public and health authorities that these rumors are just that, that no large-scale outbreak has occurred. However, resources for public health are often limited in developing or resource-limited countries. As a result, surveillance systems are sometimes not well developed and do not provide timely or useful data. To provide a common strategy for improving surveillance capacity, the World Health Organization (WHO) and other United Nations agencies have developed standardized approaches for ongoing, routine surveillance. One such system is the Integrated Disease Surveillance and Response strategy, another is the United Nations High Commissioner for Refugees Health Information System, and a third is the Early Warning and Response Network. The latter two can be activated following or during a disaster or humanitarian emergency.

21.6.1 Integrated Disease Surveillance and Response (IDSR)

In 1998, member states of the WHO African Regional Office adopted the Integrated Disease Surveillance and Response (IDSR) strategy to improve the availability and use of surveillance and laboratory data for control of communicable diseases of public health importance, reflecting their importance as leading contributors to mortality, morbidity, and disability in the African region. By 2009, 43 of the 46 countries in the region were participants in the program, which involves policy and technical support to member nations, defining minimum core capacities, training, facilitating information exchange, and improving laboratory capacity to confirm the diagnosis of priority diseases, with the final product being that of an integrated national surveillance system (World Health Organization Regional Office for Africa 2009).



Integrated Disease Surveillance Response

Morbidity (disease) and Mortality (death) Reporting Form

Gobolka *(Region)*:............................ Tuulo/Deggaan *(Village/Settlement)*:.. Epidemiological Week:..................

Degmada *(District)*:.......................... Goobta Caafimaadka *(Health Facility)*:... Laga billabo Sabtida *(Monday to Sunday)*

Hay'adda Magaca *(Supporting organization/NGO)*:... From: ..../.......... /20 ....

Magaca Qofka buuxiyey Foomka *(Name of person filling the form)*:.. Ilaa Axadda *(To)*: ..../.......... /20 ....

Lambarka telefoonka/e-mailka howlwadeenka *(Phone/e-mail)*:.. Date: ..../.......... /20 ....

Saxiixa *(Signature of person filling the form)*:...................................

	Cudurrada / Dhacdooyinka la sahminaayo *(Health Events Under Surveillance)*	*Xaalad kastoo caafimaad ukala qaybi tirada:* *(For each Health event, enter total number of: male/female)*		Dadka jirran *(Cases)*		Dhimashada *(Deaths)*	
		RAG	**DUMAR**	**< 5 yrs**	**≥ 5 yrs**	**< 5 yrs**	**≥ 5 yrs**
1	**Shuban Biyoodka Degdegga** *(Acute Watery Diarrhoea)*						
2	**Shuban Dhiig** *(Bloody Diarrhoea)*						
3	***Caabuqa Neefmareeka Daran*** ***(Oofwareen, Pneumonia*** *(SARI)****)***						
4	***Hargabka ama wax lamida*** *(Influenza Like Illness, ILI)*						
5	**Jadeeco aan la hubin** *(Suspected Measles)*						
6	**Cudurka Maskax Garaadka aan la hubin** *(Suspected Meningitis)*						
7	**Cudurka dabaysha** *(Acute Flaccid Paralysis)*						
8	**Qandhada Dhiig Baxa degdegga** *(Suspected Hemorrhagic Fever)*						
9	***Cagaarshow/Indhacaseeye degdegga*** *(Acute joundice syndrome)*						
10	**Gawracato** *(Diphtheria)*						
11	**Xiiqdheerta** *(Whooping Cough)*						
12	**Teetanda*** *(Tetanus)*			≤28 days:	>28 days:	≤28 days:	>28 days:
13	**Duumo aan la caddeyn** *(Suspected Malaria)*						
14	**Dummo la-xaqiijiyey** *(Confirmed Malaria)*						
15	**Cudurka Leyshmaniyada aan la hubin** *(Suspected Leishmania)*						
16	**Cudurka kaadi Dhiigga** *(Urinary Schistosomasis)*						
	Wadar *(TOTAL consultations)*						

NB:

- Please include only those cases that were seen at the health facility during the surveillance week. Each case should be counted only once.
- Write "0" (zero) if you had no case on death of one of the Health Events listed in the form
- For Tetanus, all cases above 28 days will be classified as adult tetanus (fill in the > 28 days /5 years column)

Fig. 21.2 Weekly integrated diseases surveillance and response reporting form used in Somalia

For example, the July 2010 IDSR Epidemiological Report provides country-specific information on trends for polio, influenza, and yellow fever and on outbreaks of cholera, meningitis, and dengue (World Health Organization Regional Office for Africa 2010). The overall objective is to identify communicable diseases and thereby reduce their burden. Data are collected at healthcare facilities on a weekly basis, the number of variables is usually less than 20. One sample form for IDSR that was used briefly in Somalia can be found in Fig. 21.2.

However, many challenges still exist for the member nations, including a national coverage, integration of the 2005 International Health Regulations, and mobilizing resources and staff, including laboratory capacity to mention a few. As with many of these surveillance systems, high staff turnover and attrition at all levels of the healthcare system impact on the quality of the data and usefulness of the system.

21.6.2 United Nations High Commissioner for Refugees Health Information System

In protracted refugee camp settings, a mix of factors contribute to disrupted collection and flow of health information, including evolving operational environments, uncertain access to the camps, uncertain communications, and staff turnover.

In addition, multiple organizations might be involved in providing camp health services and may simultaneously operate independent information systems to meet the reporting requirements of their respective organizations, donors, and host ministries of health. The result is fragmented and often incomplete and inaccurate reporting and a lack of comparability of health data across camps. To address these problems, from 1999 through 2005, the US CDC collaborated with the Office of the United Nations High Commissioner for Refugees (UNHCR) to develop a Health Information System (HIS) for refugee camps in Pakistan and Tanzania. Since 2005, the HIS has been implemented in 17 NGO operations covering 1.5 million refugees in 85 camps in Africa, Asia, and the Middle East. Its purpose was to design, monitor, and evaluate healthcare programs and thereby improve the health status of refugees under the protection of UNHCR through use of evidence to base policy formation and changes and manage the healthcare programs. Despite the use of common HIS tools and guidelines, initial evaluations demonstrated weaknesses in implementing the system and disparities in the ability of users to interpret and apply the data to improve camp health services (Haskew et al. 2010).

21.6.3 World Health Organization Early Warning and Response Network (EWARN)

During crises, preexisting disease surveillance systems may be insufficient to provide early warning of outbreaks and provide ongoing situational awareness, and special surveillance efforts might be required. Such surveillance should cover as much of the affected population as security and logistic limitations allow, and it should provide quantitative and qualitative information related to diseases with high mortality and, depending on the type of event, injuries. While involving as many stakeholders as possible, it should have a single lead agency, such as the Ministry of Health in the countries where the crisis has occurred or, if the Ministry is unable to take the leadership role, the WHO. The Early Warning and Response Network (EWARN) of the WHO's Health Action in Crises program provides guidance for surveillance during crises (World Health Organization 2011c). EWARN guidance recommends that:

- A monitoring team be established at the managerial level of the overall emergency relief operation.
- A limited number of diseases and indicators of the functionality of healthcare services be targeted for monitoring, based on priorities identified by a rapid assessment (see Sect. 21.5).
- Only experienced epidemiologists and public health officials take the lead in choosing these diseases and other indicators.
- Syndrome-based case definitions be used for surveillance, presuming that routine access to laboratory testing to confirm diagnoses is unavailable. To the extent that diagnostic tests are available at reference laboratories, syndrome-based surveillance should be supplemented by laboratory testing to confirm diagnoses in a sample of affected people, where possible.

Table 21.4 Ten communicable diseases/syndromes under surveillance, Darfur 2004 (Centers of Disease Control and Prevention, field observation, October 2009)

Disease/syndrome	Threshold to respond[a]
Acute watery diarrhea	5 cases in aged 5 or more years
Bloody diarrhea	5 cases
Acute flaccid paralysis	1 case
Suspected measles	1 case
Acute jaundice syndrome	5 cases or 1.5 times baseline
Suspected meningitis	5 cases or 1.5 times baseline
Suspected malaria	1.5 times baseline
Acute unknown fever	1.5 times baseline
Acute respiratory infection	Not defined
Neonatal tetanus	1 case

[a] Adapted from World Health Organization. Communicable disease control in emergencies: a field manual (Connolly 2005)

The EWARN approach has been used since 2004 in the Darfur crisis in Sudan, where it tracks ten communicable diseases or syndromes (Table 21.4) and where it supported detection of two disease outbreaks, one of shigellosis and another of Hepatitis E virus, and helped confirm a decline in measles incidence after a measles vaccination campaign (Musani et al. 2004). Although the EWARN approach has shown its worth, there remains a need to improve data quality, determine standard performance indicators, and improve training (World Health Organization 2010b). The threshold levels are used to initiate a response to the area where the notification was generated, which can be as little as a telephone call to verify that the report was real to as extensive as a team to investigate the potential case, determine an outbreak is occurring, and establish interventions, such as setting up cholera treatment centers or supplemental immunization activities for measles. Although EWARN was designed for crisis situations, it was recently implemented nationally in Macedonia, where communicable disease laws mandate reporting for 47 pathogens or diseases based on laboratory confirmation but where resources are lacking to provide laboratory confirmation for all diseases/pathogens on this list. Following a pilot in three cities, the country implemented the EWARN approach to monitoring eight syndrome categories of disease (Kisman et al. 2010).

21.6.4 Examples of Specific Developing-Country Uses: Haiti 2010

National Sentinel Site Surveillance Following the earthquake in Haiti in January 2010, two surveillance systems were established through a joint effort of the Haitian Ministry of Public Health and Population, PAHO, and US CDC (Centers for Disease Control and Prevention 2010d). The first, the National Sentinel

Site Surveillance (NSSS) system, was established within 2 weeks of the earthquake because there was no existing surveillance system that could provide timely data on disease trends, detect outbreaks, or characterize the affected population to target relief efforts (Centers for Disease Control and Prevention 2010d). NSSS engaged 51 hospital and clinic sites that had previously been affiliated with the US President's Emergency Plan for AIDS Relief (PEPFAR) and involved daily reporting by e-mail or telephone for 25 specified reportable conditions. During the first 3 months, surveillance data were accumulated for over 42,000 persons with reportable conditions. Nationally, the four most frequently reported conditions were acute respiratory infection (16%), injuries (12%), suspected malaria (10%), and fever of unknown cause (10%).

Internally Displaced Persons Surveillance System In just over 1 month following the earthquake, the Internally Displaced Persons Surveillance System (IDPSS) was instituted because of the increased risk of outbreaks among the more than one million displaced persons living in the hundreds of camps that spontaneously sprang up in and around Port-au-Prince (Centers for Disease Control and Prevention 2010b). This system allowed sharing and tracking of illness data among the NGOs that supported the camps and involved monitoring of six immediately notifiable conditions (suspected cases of acute hemorrhagic fever, measles, rabies, meningococcal meningitis, acute flaccid paralysis, or diphtheria), suspected cases of ten additional communicable diseases with less risk of outbreak potential or mortality, and persons with either of the three conditions considered as "programmatic indicators" (tuberculosis with interrupted treatment, HIV/AIDS with interrupted antiretroviral therapy, and pregnancy in any trimester with complications or third trimester pregnancy in a woman who had not received prenatal care).

At the time the IDPSS was established, the risk of cholera was believed to be low since cholera outbreaks had not been seen in Haiti for at least a century (Centers for Disease Control and Prevention 2010c). For the first 5 weeks of the camp surveillance, 31 NGOs reported over 96,000 new clinic visits of which 23,000 were by persons with one of the 19 reportable conditions. The three most common conditions were acute respiratory infection (9%), suspected malaria (5%), and watery diarrhea/not suspected cholera (5%) (Centers for Disease Control and Prevention 2010b).

Many previously documented challenges of postdisaster public health surveillance were experienced in Haiti, including logistical constraints, absence of baseline information on disease trends, unavailable population data to define disease rates, underreporting, lack of clarity or consistency in using surveillance case definitions, and limited laboratory capacity. Despite these challenges, both systems were valuable elements of the public health response, providing daily or weekly reports to public health partners in Haiti during the emergency response and serving as tools to respond to rumors or concerns of increases in disease (Centers for Disease Control and Prevention 2010b, d).

21.7 Disaster and Crisis Surveillance in Developed Countries

Although natural and other disasters can affect developed countries with devastating consequences, for a comparable geologic or atmospheric event, the human health impact is likely to be less severe in developed than less developed countries for multiple reasons, including a greater likelihood that homes or buildings can withstand physical stresses, stronger infrastructures to support essential services, more stable governments, and greater healthcare capacity. Moreover, developed countries have not been directly affected by large-scale complex humanitarian emergencies, such as those described elsewhere in this chapter. While the questions likely to be asked of surveillance systems during emergency or urgent situations are comparable in developed and less developed countries, preexisting surveillance capacities are generally greater in developed countries. This difference affects the relative dependence on enhancements or modifications to existing surveillance systems versus establishing event-specific surveillance systems. Despite multiple advantages of developed countries, experience with recent disasters highlights vulnerabilities in their surveillance capacity. This section presents examples of event-specific and disaster-directed surveillance in the United States and Europe as well as steps taken in the United States following Hurricane Katrina to improve uniformity of disaster surveillance tools.

21.7.1 Examples of Event- and Disaster-Related Surveillance Uses: United States

Role of Automated Syndromic Surveillance Investments in the United States in automated surveillance systems aimed at the early detection and prompt characterization of outbreaks were prompted largely by concerns in the early 2000s about the threat of potential large-scale bioterrorist attacks, and the threat of bioterrorism persists (National Intelligence Council 2004). However, in the absence of bioterrorist attacks, epidemiologists have found additional uses for these new surveillance systems and have embraced an "all-hazards" approach in their use (Buehler et al. 2008, 2009). Their ability to provide early detection of a potentially devastating bioterrorist attack has not been tested in practice, but their performance in detecting more common types of disease outbreaks has been mixed, depending on the nature of outbreaks, the type of healthcare services that affected people use, and the approach employed to conduct surveillance in specific localities. Based on observations from actual outbreaks and mathematical models, in general, the larger the outbreak and the more rapid the upswing in numbers of affected people, the more likely that syndromic approaches will provide an early indication of the presence of an outbreak (Buckeridge et al. 2005). Reflecting on the substantial impacts of influenza on seasonal patterns of respiratory disease, monitoring the impact of influenza-like illness has proved to be one of the most common uses of syndromic surveillance (Olson et al. 2007; Buehler et al. 2008), and syndromic surveillance systems became part of the mosaic of approaches used to monitor the emergence and course of H1N1 pandemic influenza in 2009–2010. In situations where environmental exposures to airborne irritants are widespread, such as exposures to smoke arising from massive wildfires or airborne particulates arising from a volcanic eruption, syndromic surveillance has proved useful in assessing the population-level impact on respiratory disease (Johnson et al. 2005; Elliot et al. 2010). Although it is difficult to define the capacity of syndromic surveillance to exclude the presence of an outbreak, when combined with other sources of information, it can provide an important "piece of the puzzle" in providing assurance that outbreaks are not occurring when the questions about threats to public health are raised (Buehler et al. 2009).

Hurricane Katrina In August 2005, Hurricane Katrina not only affected coastal regions of the United States along the Gulf of Mexico but also resulted in large numbers of evacuees migrating to other states. Because of the disruption of public health services in the New Orleans region, an ad hoc surveillance system was established that depended largely on manual data collection for a limited number of infectious diseases, non-infectious conditions, and injuries, based on reporting from selected healthcare facilities in the area for a time-limited period. Reporting from these sites was supplemented by event-specific reporting from evacuation centers and coroners. Despite the extensive structural damage and flooding in affected coastal areas and the large number of displaced persons, no major outbreaks of disease or hazardous environmental exposures were detected (Centers for Disease Control and Prevention 2005). In Georgia, where a substantial number of evacuees migrated, preexisting automated syndromic surveillance was used during the event to track emergency department use and retrospectively to inform planning for future hurricanes (Cookson et al. 2008). More broadly, US CDC sought to compile surveillance data from twelve states that were either directly affected by the hurricane or were the destination for evacuees from coastal areas. However, the ability to compile a multistate profile of the health impacts of the hurricane, especially among displaced people temporarily residing in evacuation centers, was limited by variations in surveillance methods and approaches among states (Centers for Disease Control and Prevention 2006). Following Hurricane Katrina, recognition of this variability prompted the US CDC to work with state and local public health authorities to develop, revise, and pilot tools for shelter health assessments and for morbidity and mortality surveillance after natural disasters (Schnall et al. 2011). These tools are now available on the US CDC website (Centers for Disease Control and Prevention 2011). They include a shelter assessment tool to assist environmental health practitioners to conduct a rapid assessment of shelter conditions and make recommendations for improvements, a mortality surveillance form to identify number and cause of deaths related to disasters, an individual morbidity surveillance form to capture individual-level medical conditions treated in shelters or potentially hospitals during a disaster, a morbidity surveillance line list form to summarize or collect less detailed data on medical conditions of multiple individuals seen in shelters or hospitals, and tally and summarize sheets to aggregate morbidity data from individual or line list forms.

Population Health Surveys Mental health impacts of disasters are well recognized, mental, and physical effects of disasters may be interactive, and needs for mental health services may be substantial. Following the September 2001 terrorist attacks on the World Trade Center in New York City, a telephone survey using a random-digit dialing sample design targeting households in the borough where the World Trade Center was located. Questions addressed respondents' demographic characteristics, measures of exposure to the event (e.g., directly witnessing the event, having a friend or relative killed in the attack), measures of social support or resiliency (e.g., social support networks), and mental health status (e.g., sets of questions that reflect the presence of post-traumatic stress disorder (PTSD) or depression symptoms). Survey findings demonstrated a "substantial burden" of PTSD and depression symptoms, even though the post-event approach to implementing the survey precluded clear pre- and post-event comparisons (Galea et al. 2002). In addition to ongoing monitoring of trends in health-related behaviors that might be affected by emotional distress, health departments of New York and two neighboring states supplemented an existing state-based population health survey, the Behavioral Risk Factor Surveillance System, to assess psychological and emotional effects of the attacks, such as anger, nervousness, and sleep problems (Centers for Disease Control and Prevention 2002b). Multiple surveys examining the affected populations of disasters and conflicts have also been performed with persons in less developed countries and similar results are found (Sabin et al. 2003; Cardozo et al. 2004; Thienkrua et al. 2006; van Griensven et al. 2006). The majority of these surveys use population-based sampling and standardized instruments, such as the Hopkins Symptom Checklist-25 (Mollica et al. 1987) that identifies elevated levels of symptoms of anxiety and depression and the Harvard Trauma Questionnaire (Mollica et al. 1992) that measures trauma events and PTSD symptoms.

21.8 Conclusions

Globally, there is an increasing threat of disasters and accompanying health impacts due to multiple factors: population growth, the increasing concentration of populations in urban areas due to migration, the location of many large urban centers in coastal areas, and the impact of climate change on rising sea levels and the increasing threat of extreme weather events (Freeman et al. 2003; James et al. 2008). The health impacts of natural disasters on affected populations are likely to differ in developed and less developed countries, where high concentrations of poverty and lack of infrastructure in the latter will result in heightened vulnerability to the adverse effects of disasters and where weak governance and marginalized population conditions are more likely to prompt humanitarian crises. Despite these differences between developed and less developed countries, when disasters and humanitarian emergencies occur, effective surveillance systems will be needed by public health response agencies to detect and track outbreaks, identify and monitor health status and urgent healthcare needs, and inform efforts to minimize adverse health effects.

The key components and requirements of surveillance systems in disasters and humanitarian emergencies are to only capture those conditions or diseases that have the greatest potential of causing high morbidity and mortality, that if not controlled can lead to epidemics and further morbidity and mortality, and for which an intervention exists in the area affected or through advocacy can be brought into the area. Therefore, the system needs to be simple, capturing as few conditions or diseases as possible; flexible to allow for addition of diseases if new pathogens are identified, such as cholera in October 2010 in Haiti; and acceptable to all partners so participation is high and completeness of reporting maintained; the data flow must be well characterized and understood so data can be collected, reported, and analyzed in a timely fashion. Finally, the information must be valid (requiring continuous training and frequent monitoring), useful, and result in action – data should be collected for the sole purpose of detecting and responding to diseases among the affected population.

References

Brammer L, Blanton L, Epperson S, Mustaquim D, Bishop A, Kniss K, Dhara R, Nowell M, Kamimoto L, Finelli L (2011) Surveillance for influenza during the 2009 influenza A (H1N1) pandemic–United States, April 2009–March 2010. Clin Infect Dis 52(Suppl 1):S27–S35

Buckeridge DL, Burkom H, Campbell M, Hogan WR, Moore AW (2005) Algorithms for rapid outbreak detection: a research synthesis. J Biomed Inform 38(2):99–113

Buehler JW, Berkelman RL, Hartley DM, Peters CJ (2003) Syndromic surveillance and bioterrorism-related epidemics [see comment]. Emerg Infect Dis 9(10):1197–1204

Buehler JW, Sonricker A, Paladini M, Soper P, Mostashari F (2008) Syndromic surveillance practice in the United States: findings from a survey of state, territorial, and selected local health departments. Adv Dis Surveill 6(3):1–8

Buehler JW, Whitney EA, Smith D, Prietula MJ, Stanton SH, Isakov AP (2009) Situational uses of syndromic surveillance. Biosecur Bioterror 7(2):165–177

Camp Coordination Camp Management Cluster in Haiti (2011) DTM data. http://www.cccmhaiti.info/. Accessed on 14 Sept 2011

Cardozo BL, Talleya L, Burton A, Crawford C (2004) Karenni refugees living in Thai–Burmese border camps: traumatic experiences, mental health outcomes, and social functioning. Soc Sci Med 58:2637–2644

Centers for Disease Control and Prevention (2002a) Syndromic surveillance for bioterrorism following the attacks on the World Trade Center – New York City, 2001. Morb Mortal Wkly Rep 51(Special issue):13–15

Centers for Disease Control and Prevention (2002b) Psychological and emotional effects of the September 11 attacks on the World Trade Center – Connecticut, New Jersey, and New York, 2001. Morb Mortal Wkly Rep 51(35):784–786

Centers for Disease Control and Prevention (2004) Introduction to investigating an outbreak. http://www.cdc.gov/excite/classroom/outbreak/objectives.htm. Accessed 10 Apr 2011

Centers for Disease Control and Prevention (2005) Surveillance for illness and injury after Hurricane Katrina – New Orleans, Louisiana, September 8–25, 2005. Morb Mortal Wkly Rep 54(40):1018–1021

Centers for Disease Control and Prevention (2006) Morbidity surveillance after Hurricane Katrina – Arkansas, Louisiana, Mississippi, and Texas, September 2005. Morb Mortal Wkly Rep 55(26):727–731

Centers for Disease Control and Prevention (2009) Community Assessment for Public Health Emergency Response (CASPER) Toolkit. http://emergency.cdc.gov/disasters/surveillance/pdf/CASPER_toolkit_508%20COMPLIANT.pdf. Accessed 10 Apr 2011

Centers for Disease Control and Prevention (2010a). Early aberration reporting system (EARS). http://emergency.cdc.gov/surveillance/ears/. Accessed 10 Apr 2011

Centers for Disease Control and Prevention (2010b) Rapid establishment of an internally displaced persons disease surveillance system after an earthquake – Haiti, 2010. Morbid Mortal Wkly Rep 59(30):939–945

Centers for Disease Control and Prevention (2010c) Update: outbreak of cholera – Haiti, 2010. Morbid Mortal Wkly Rep 59(48):1586–1590

Centers for Disease Control and Prevention (2010d) Launching a national surveillance system after an earthquake – Haiti, 2010. Morbid Mortal Wkly Rep 59(30):933–938

Centers for Disease Control and Prevention (2011) Public health assessment and surveillance after a disaster. http://www.bt.cdc.gov/disasters/surveillance/. Accessed 24 Apr 2011

Checchi F, Roberts L (2005) Humanitarian practice network report No 52. Interpreting and using mortality data in humanitarian emergencies: a primer for non-epidemiologists. http://www.odihpn.org/documents/networkpaper052.pdf. Accessed 10 Apr 2011

Connolly MA (ed) (2005) Communicable disease control in emergencies: a field manual. World Health Organization, Geneva

Cookson ST, Soetebier K, Murray EL, Fajardo GC, Hanzlick R, Cowell A, Drenzek C (2008) Internet-based morbidity and mortality surveillance among Hurricane Katrina evacuees in Georgia. Prev Chronic Dis 5(4):1–7

Elliot AJ, Singh N, Loveridge P, Harcourt S, Smith S, Pnaiser R, Kavanagh K, Robertson C, Ramsay CN, McMenamin J, Kibble A, Murray V, Ibbotson S, Catchpole M, McCloskey B, Smith GE (2010) Syndromic surveillance to assess the potential public health impact of the Icelandic volcanic ash plume across the United Kingdom, April 2010. Eurosurveillance 15(23):Article 3

Federal Emergency Management Agency (2007) National incident management system [FEMA 501/Draft August 2007]. http://www.fema.gov/library/viewRecord.do?id=2961. Accessed 14 Apr 2011

Foege WH, Hogan RC, Newton LH (1976) Surveillance projects for selected diseases. Int J Epidemiol 5(1):29–37

Freeman PK, Keen M, Mani M (2003) Dealing with increased risk of natural disasters: challenges and options, IMF Working Paper WP/03/197. http://www.imf.org/external/pubs/cat/longres.cfm?sk=16830.0. Accessed 24 Apr 2011

Galea S, Ahern J, Resnick H, Kilpatrick D, Bucuvalas M, Gold J, Vlahov D (2002) Psychological sequeale of the Septermber 11 terrorist attacks in New York City. N Engl J Med 346(13):982–987

Garfield R, Blake C, Chatainger P, Walton-Ellery S (2011) Humanitarian practice network report no. 69. Common needs assessments and humanitarian action. http://www.humanitarianforum.org/data/files/resources/807/en/commonneedsassessment.pdf. Accessed 20 Apr 2011

Gerrity ET, Flynn BW (1997) Mental health consequences of disasters (chapter 6). In: Noji EK (ed) The public health consequences of disasters. Oxford University Press, New York, pp 101–132

Green DC, Buehler JW, Silk BJ, Thompson NJ, Schild LA, Klein M, Berkelman RL (2006) Trends in healthcare use in the New York City region following the terrorist attacks of 2001. Biosecur Bioterror 4(3):263–275

Haskew C, Spiegel P, Tomczyk B, Cornier N, Hering H (2010) A standardized health information system for refugee settings: rationale, challenges and the way forward. Bull World Health Organ 88(10):792–794

Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D (2004) Syndromic surveillance in public health practice, New York City. Emerg Infect Dis 10(5):858–864

Inter-Agency Standing Committee (2009) Inter-agency standing committee. http://www.humanitarianinfo.org/iasc/. Accessed 10 Apr 2011

James JJ, Subbarao I, Lanier WL (2008) Improving the art and science of disaster medicine and public health preparedness. Mayo Clin Proc 83(5):559–562

Johnson JM, Hicks L, McClean C, Ginsberg M (2005) Leveraging syndromic surveillance during the San Diego wildfires, 2003. In: Syndromic surveillance: reports from a national conference, 2004. Morbid Mortal Wkly Rep 54(Suppl):190

Kisman M, Donev D, Kisman A (2010) International standards and strategies for the surveillance, prevention and control of brucellosis. Maced J Med Sci 3(3):273–277

Lombardo JS, Sniegoski CA, Loschen WA, Westercamp M, Wade M, Dearth S, Zhang G (2008) Public health surveillance for mass gatherings. Johns Hopkins APL Tech Dig 27(4):347–355

Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani P, Mostashari F, Pavlin JA, Gesteland PH, Treadwell T, Koski E, Hutwagner L, Buckeridge DL, Aller RD, Grannis S (2004) Implementing syndromic surveillance: a practical guide informed by the early experience. J Am Med Inform Assoc 11(2):141–150

Médecins Sans Frontières (1997) Refugee health: an approach to emergency situations. Macmillan Education, Oxford

Mokdad AH, Mensah GA, Posner SF, Reed E, Simoes EJ, Engelgau MM, Chronic Diseases and Vulnerable Populations in Natural Disasters Working Group (2005) When chronic conditions become acute: prevention and control of chronic diseases and adverse health outcomes during natural disasters. Prev Chronic Dis 2 (Special Issue):1–4

Mollica RF, Caspi-Yavin Y, Marneffe D, Khuon F, Lavelle J (1987) Indochinese versions of the Hopkins Symptom Cecklist-25: a screening instrument for the psychiatric care of refugees. Am J Psychiatry 144:497500

Mollica RF, Wyshak G, Bollini P, Truong T, Tor S, Lavelle J (1992) The Harvard Trauma Questionnaire. Validating a cross-cultural instrument for measuring torture, trauma, and posttraumatic stress disorder in Indochinese refugees. J Nerv Ment Dis 180:111116

Musani A, Sabatinelli G, Koller T, Nabarro D (2004) The challenges of securing health in humanitarian crises. Bull World Health Organ 82(9):642

National Intelligence Council (2004) Mapping the global future: report of the National Intelligence Council's 2020 Project. N. I. Council. Government Printing Office, Pittsburgh

Olson DR, Heffernan RT, Paladini M, Konty K, Weiss D, Mostashari F (2007) Monitoring the impact of influenza by age: emergency department fever and respiratory complaint surveillance in New York City. PLoS Med 4(8):e247

OneResponse (2013) http://oneresponse.info. Accessed 13 Sept 2011

Pan American Health Organization (1981) Emergency health management after natural disaster, Scientific Publication No. 407. Pan American Health Organization, Washington, DC

Pan American Health Organization (2010) Health cluster bulletin cholera and post-earthquake response in Haiti – #20. http://www.reliefweb.int/rw/rwb.nsf/db900SID/MUMA-8EJ2KK?OpenDocument&rc=2&emid=EP-2010--000210-HTI. Accessed 11 Apr 2011

reliefweb (2013) http://www.reliefweb.int. Accessed 13 Sept 2011

Sabin M, Lopes Cardozo B, Nackerud L, Kaiser R, Varese L (2003) Factors associated with poor mental health among Guatemalan refugees living in Mexico 20 years after civil conflict. J Am Med Assoc 290(5):635–642

Schnall AH, Wolkin AF, Noe R, Hausman LB, Wiersma P, Soetebier K, Cookson ST (2011) Evaluation of a standardized morbidity surveillance form for use during disasters caused by natural hazards. Prehosp Disaster Med 26(2):90–98

Spiegel PB, Le P, Maloney S, van der Veen A (2004) Quality of malnutrition assessment surveys conducted during famine in Ethiopia. J Am Med Assoc 292(5):613–618

Spiegel PB, Salama R, Ververs M-T, Salama P (2007) Occurrence and overlap of natural disasters, complex emergencies and epidemics during the past decade (1995–2004). Confl Health 1(2). doi:10.1186/1752–1505–1–2

Standardized Monitoring and Assessment of Relief and Transitions (2009) SMART (Standardized Monitoring and Assessment of Relief and Transitions). http://www.smartmethodology.org/. Accessed 10 Apr 2011

Stergarchis A, Tsouros AD (2007) Overview and framework (chapter 1). In: Tsouros AD, Efstathiou PA (eds) Mass gatherings and public health: the experience of the Athens 2004 Olympic Games. World Health Organization Regional Office for Europe, Copenhagen, pp 3–28

Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV, CDC PulseNet Task Force (2001) PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerg Infect Dis 7(3):382–389

Swerdlow DL, Finelli L, Bridges CB (2011) 2009 H1N1 Influenza pandemic: field and epidemiologic investigations in the United States at the start of the first pandemic of the 21st century. Clin Infect Dis 52(Suppl 1):S1–S3

Tan CG, Sandhu HS, Crawford DC, Redd SC, Beach MJ, Buehler JW, Bresnitz EA, Pinner RW, Bell BP, Regional Anthrax Surveillance T; Centers for Disease Control and Prevention New Jersey Anthrax Surveillance Team (2002) Surveillance for anthrax cases associated with contaminated letters, New Jersey, Delaware, and Pennsylvania, 2001. Emerg Infect Dis 8(10):1073–1077

Thienkrua W, Cardozo BL, Chakkraband MLS, Guadamuz TE, Pengjuntr W, Tantipiwatanaskul P, Sakornsatian S, Ekassawin S, Panyayong B, Varangrat A, Tappero JW, Schreiber M, van Griensven F, The Thailand Post-Tsunami Mental Health Study Group (2006) Symptoms of posttraumatic stress disorder and depression among children in tsunami-affected areas in southern Thailand. J Am Med Assoc 296(5):549–559

Toole MJ, Waldman RJ (1990) Prevention of excess mortality in refugee and displaced populations in developing countries. J Am Med Assoc 263(24):3296–3302

Traeger MS, Wiersma ST, Rosenstein NE, Malecki JM, Shepard CW, Raghunathan PL, Pillai SP, Popovic T, Quinn CP, Meyer RF, Zaki SR, Kumar S, Bruce SM, Sejvar JJ, Dull PM, Tierney BC, Jones JD, Perkins BA, Florida Investigation Team (2002) First case of bioterrorism-related inhalational anthrax in the United States, Palm Beach County, Florida, 2001. Emerg Infect Dis 8(10):1029–1034

United Nations Country Team in Pakistan (2005) Acute diarrhoea outbreaks in spontaneous camps – more resources needed to support Government of Pakistan. http://www.reliefweb.int/rw/rwb.nsf/db900sid/EGUA-6HYPCX?OpenDocument. Accessed 10 Apr 2011

United Nations Radio (2010) 2010 deadliest year for natural disasters http://www.unmultimedia.org/radio/english/detail/112156.html. Accessed 10 Apr 2011

van Griensven F, Chakkraband MLS, Thienkrua W, Pengjuntr W, Lopes Cardozo B, Tantipiwatanaskul P, Mock PA, Ekassawin S, Varangrat A, Gotway C, Sabin M, Tappero JW, The Thailand Post-Tsunami Mental Health Study Group (2006) Mental health problems among adults in Tsunami-affected areas in Southern Thailand. J Am Med Assoc 296(5):537–548

Watson JT, Gayer M, Connolly MA (2007) Epidemics after natural disasters. Emerg Infect Dis 13(1):1–5

World Health Organization (2006) Guidance note on using the cluster approach to strengthen humanitarian response. http://www.who.int/hac/network/interagency/news/cluster_approach/en/. Accessed 10 Apr 2011

World Health Organization (2010a) Initial rapid assessment (IRA) of flood affected districts in Punjab – August 2010. http://www.whopak.org/idps/documents/assessments/IRA-Assessment-Report%20in%20Punjab.pdf. Accessed 10 Apr 2011

World Health Organization (2010b) Early warning surveillance and response in emergencies, Report of the WHO technical workshop, 7–8 December 2009. http://whqlibdoc.who.int/hq/2010/WHO_HSE_GAR_DCE_2010.4_eng.pdf. Accessed 24 Apr 2011

World Health Organization (2011a) Health action in crises: definitions: emergencies. http://www.who.int/hac/about/definitions/en/index.html. Accessed 13 Sept 2011

World Health Organization (2011b) Health action in crises: cluster countries. http://www.who.int/hac/global_health_cluster/countries/en/. Accessed 11 Apr 2011

World Health Organization (2011c) Health action in crises: assuring information on health and its determinants. http://www.who.int/hac/techguidance/tools/manuals/who_field_handbook/5/en/index2.html. Accessed 22 Apr 2011

World Health Organization Regional Office for Africa (2009) Integrated disease surveillance (IDS) overview. http://www.afro.who.int/en/clusters-a-programmes/dpc/integrated-disease-surveillance/overview.html. Accessed 11 Apr 2011

World Health Organization Regional Office for Africa (2010) Integrated disease surveillance and response (IDSR) epidemiological report, July 2010. http://www.afro.who.int/en/clusters-a-programmes / dpc/integrated-disease-surveillance/features/2473-idsr-epidemiological-report-july-2010.html. Accessed 11 April 2011

Screening 22

Anthony B. Miller

Contents

A.B. Miller
Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_32,

22.1 Introduction

22.1.1 General Principles of Screening

Screening, sometimes termed “secondary prevention,” is one of the major components of disease control, the others comprising primary prevention, diagnosis, treatment, rehabilitation after treatment or disability, and palliative care. Ideally, the control of a disease should be achievable, either by preventing the disease from occurring or, if it does occur, by curing those who develop it by appropriate treatment. Complete success from prevention would make treatment obsolete. Complete success from treatment, however, would not make prevention obsolete, as there are costs and undesirable sequelae from the disease and treatment that patients and society would like to avoid if at all possible, especially from diseases such as cancer, diabetes, and hypertension. At present, neither is completely successful for most diseases; they will continue to complement each other for a number of conditions, while screening can be regarded as complementary to one or both of the other approaches.

Because of the deep-rooted belief among physicians that “early diagnosis” of disease is beneficial, many regard screening as bound to be effective. However, for a number of reasons discussed below, this is not necessarily so, as shown by the failure of screening for lung cancer using sputum cytology or chest X-rays to reduce mortality from the disease (e.g., Prorok et al. 1984). It is the purpose of this chapter to attempt to define some of the fundamental issues that are relevant to the consideration of screening in disease control. The approach taken will be from the epidemiology, or public health viewpoint, rather from the clinical standpoint.

Although it is often assumed that screening tests must involve some sort of technological procedure, such as an X-ray or laboratory test, screening can involve simple clinical examinations, such as assessment of blood pressure or a clinical breast examination. However, it is the advent of expensive technologically based screening tests in the last few decades which has focused attention on the need for critical evaluation of screening and the importance of soundly based screening programs.

22.1.2 Definition

Screening was defined by the United States Commission on Chronic Illness (1957) as: “the presumptive identification of unrecognized disease or defect by the application of tests, examinations or other procedures that can be applied rapidly.” A screening test is not intended to be diagnostic. Rather a positive finding will have to be confirmed by special diagnostic procedures.

By definition, screening is offered to those who do not suspect that they may have a disease. This is subtly different from being asymptomatic. Symptoms may be revealed by careful questioning related to the organ of interest, not regarded by the individual attending for screening as being related to a possible disease. Further, in a public health program, open to all comers, it may not be possible to determine that all subjects who enroll are truly asymptomatic. Indeed, many may enroll because they have a suspicion that they have the disease of interest, in the hope that their suspicion will not be confirmed. Thus, although it is usual to make the assumption that participants in screening programs are asymptomatic, this is not a necessary nor an absolute prerequisite for participation in public health-based screening programs.

22.2 General Principles Governing the Introduction of Screening

The principles that should govern the introduction of screening programs were first enunciated by Wilson and Junger (1968) and refined since (e.g., by Miller 1978, 1996; Cuckle and Wald 1984; Strong et al. 2005). These will now be considered:

1. *The disease should be an important health problem*. In practical terms, this means that the disease prevalence should be high and the disease should be the cause of substantial mortality and/or morbidity. However, it is important to recognize that the life expectancy of a screened population may be changed little even if the program is successful. In most technically advanced countries, for example, even if all cancer were to be eradicated, the effect of other competing causes of death is such that life expectancy would be increased only by about 2 1/2 years. The benefits of cholesterol screening to prevent heart disease are measured, at the population level, in days. It is not possible to provide a precise estimate for the level of burden necessary to mount a screening program. The level of morbidity and mortality considered to be important will depend on a combination of factors such as the age distribution of the population affected or the severity of the illness. There may be certain circumstances when the major benefit from screening follows, not from reduction in mortality but from reduction of morbidity consequent upon the diagnosis of the disease in a more treatable phase in its natural history. This could mean that the extent of treatment required and the possibility that treatment may be debilitating or mutilating would be much less. Such advantages may be difficult to quantify; however, as they may be considerable in psychological terms to individuals, and to communities in the lowering in the requirements for extensive rehabilitation services, they should not be overlooked.
2. *The disease should have a detectable preclinical phase (DPCP)*. It is important to recognize that this principle is not as follows: "The natural history of the disease should include a phase with a detectable precursor." For example for many cancers, including breast and prostate cancer, the DPCP is largely asymptomatic invasive cancer (Fig. 22.1). For cervix cancer, on the other hand, the DPCP probably includes the whole range from dysplasia (Cervical Intraepithelial Neoplasia (CIN) 1 or Low grade Squamous Intra-epithelial Lesion (LSIL)) through to occult invasive cancer (Fig. 22.2). An alternative name given to the

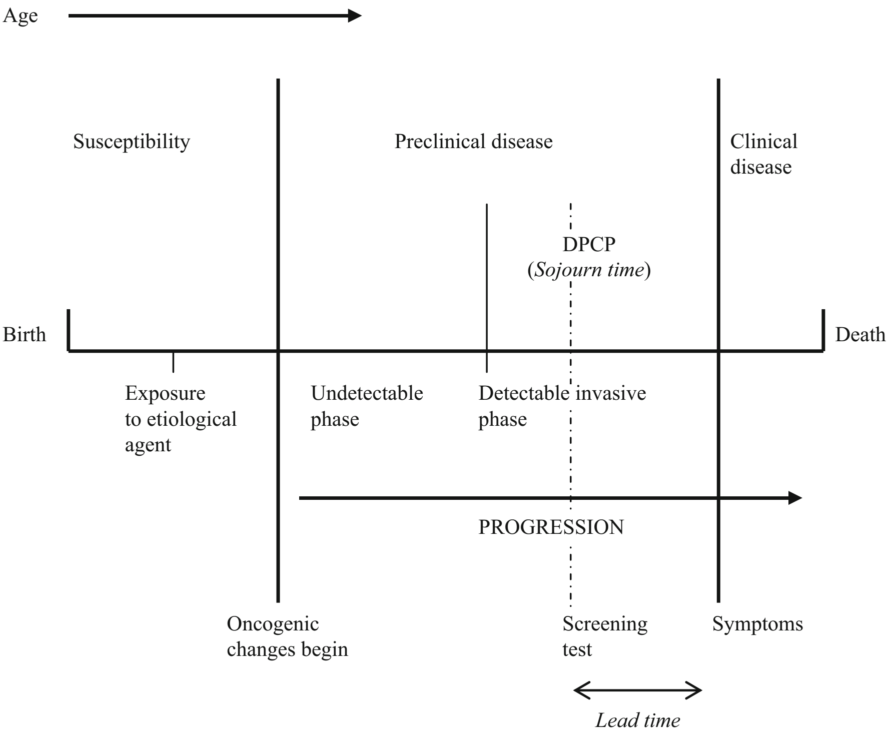


Fig. 22.1 The natural history of a cancer when the detectable preclinical phase is asymptomatic invasive disease

DPCP is sojourn time, meant to indicate the period during which a detectable lesion "sojourns" in the organ. In screening for cardiovascular disease with cholesterol or hypertension, physiological markers of risk are being identified rather than detectable precursors. Nevertheless, such a marker can be considered as synonymous to a preclinical phase.

3. *The natural history of the condition should be known*. Ideally such a requirement implies that it is known at what stage in the disease process progression, disability, and/or death can no longer be prevented. If such information was available and the stage that the development of the disease had reached in individuals was determinable, it would be possible to decide precisely when a screening test should be applied in order to achieve maximum benefit and minimal overutilization of resources. In a disease such as cervix cancer, it is now recognized that the majority of the preclinical abnormalities detected by the cytology test are not destined to progress and that the majority will regress, as depicted in Fig. 22.2 (Boyes et al. 1982; Holowaty et al. 1999). This makes it imperative that screening is planned to ensure that gross overtreatment of the majority does not occur. The same requirement may be even more necessary for

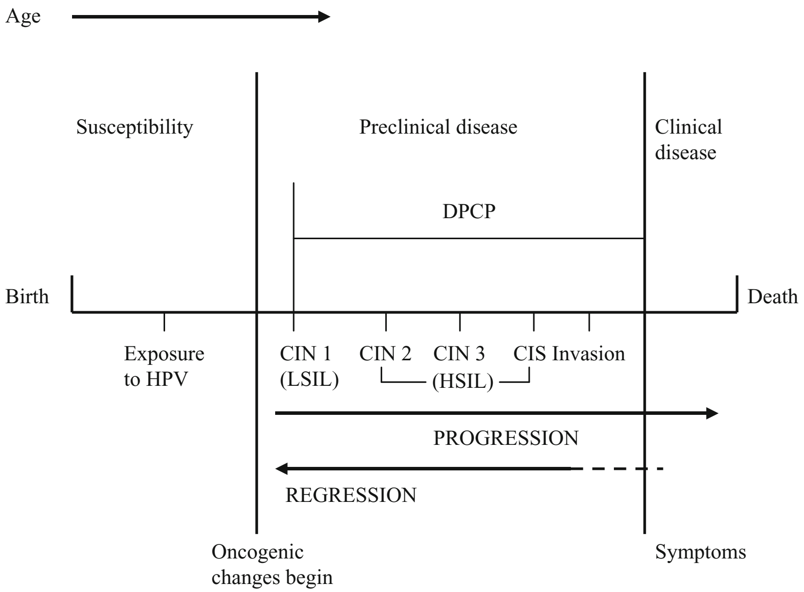


Fig. 22.2 The natural history of cancer of the cervix

the test for the primary etiological agent of the disease and the test for evidence of infection with high-risk types of the human papilloma virus (Miller 2004). Such conclusions have substantial implications with regard to the optimum frequency of screening examinations. Designing a program directed to those lesions that, in the absence of screening, will progress and more rapidly escape curability, if they can be identified, will be the appropriate approach. Designing a program which maximizes the detection of all cases, the majority with a good prognosis but which in the absence of screening may be unlikely to progress, will waste resources. Such knowledge can be applied to the population in planning screening programs, but unfortunately, it seems unlikely that knowledge will be accumulated to make it possible to determine the natural history of disease in individuals within the population sufficiently precisely. It is recognized that the rate of progression of clinically detected disease from the point of diagnosis to cure or to death varies substantially in different individuals. The distribution of rates of progression of preclinically detectable disease that might be identified by screening is likely to be equally wide. Thus, although an objective for research on screening has to determine the extent of the distribution of the sojourn times of the DPCP, in considering the introduction of screening programs and the scheduling of tests within programs, it is necessary to balance benefits with costs. This means that the schedule will have to be determined that will enable the detection of the maximal number of still curable cases compatible with the longest interval between tests, principles which unfortunately are often anathema to clinicians.

4. *The disease should be treatable, and there should be a recognized treatment for lesions identified following screening*. This principle can be elaborated as follows:

   There should be evidence of the effectiveness of treatment of lesions discovered as a result of screening in reducing disease incidence and/or mortality and the level of improvement expected should be stated,

   and secondly:

   There should be a reasonable expectation that recommendations for the appropriate management of the lesions discovered from a screening program will be complied with both by the individual with the lesion and by the physician responsible for his (or her) health care.

   Screening programs should only be set up when there are adequate facilities for treating lesions discovered as a result of screening and functioning referral systems for securing such treatment. There is obviously no point in establishing a screening program and identifying lesions that should be treated if the facilities, or the infrastructure, are not available for referral, confirmation of diagnosis, and treatment. In general, this is not a problem for technically advanced countries, but it can be for developing countries. Unfortunately problems allied to these have occurred. Thus, on occasions, it has not been certain whether or not lesions identified as a result of screening should be regarded as true disease precursors. When lesions are first identified in a screening program, information may not be available as to their appropriate treatment, and special studies may be required. Otherwise, errors in terms of observation rather than treatment on the one hand or too extensive treatment on the other are possible. In prostate cancer screening, for example, if too radical treatment is applied in the elderly to the latent or good prognosis prostate cancers that may be identified in a screening program, the morbidity in terms of incontinence and impotence, and even the mortality from treatment, could offset any benefit from the earlier detection of lesions with truly malignant potential (Chodak and Schoenberg 1989; Miller 1991; Krahn et al. 1994; Andriole et al. 2009).

   A different sort of difficulty could arise when, as a result of screening, lesions are diagnosed earlier in their natural history, but in spite of this, death is still inevitable. For example, if the available screening methods will not succeed in diagnosing disease before it is outside the range of current therapy, then screening to detect such disease is not worthwhile. The early studies of screening for lung cancer suggested this was not a condition amenable to screening, probably for this reason (Prorok et al. 1984). Recently, evidence has accrued from uncontrolled studies that low-dose helical (spiral) computerized tomography of the lung is capable of detecting approximately four times as many small stage 1 lung cancers as chest X-rays and that these appear likely to have a good prognosis (e.g., Henschke et al. 1999). However, these are peripheral cancers and are largely adenocarcinomas or related cancers so that the technique may not be capable of detecting early either the majority of squamous cell or of small cell cancers and thus may be selectively detecting only those cancers with a good prognosis. If, after spiral CT scanning, the majority of the lung cancers with a poor prognosis continue to occur at about the same time in their natural history as at the present, there could be little overall effect upon lung cancer mortality in the spiral CT screened group (Bach et al. 2007). Yet considerable costs will be incurred in treating the peripheral lesions not destined to progress. It is for this reason that large-scale trials of this approach are under way in the United States and Europe.

5. *The screening test to be used should be acceptable and safe*.
   In general, this implies a non-invasive test with high validity. Other criteria of a good screening test include ease of use and relatively low cost. These principles and approaches to assessing validity will now be discussed.

22.2.1 The Validity of a Screening Test

Two measures suffice to describe the validity of screening tests: sensitivity and specificity.

Sensitivity is defined as the ability of a test to detect all those with the disease in the screened population. This is expressed as the proportion of those with the disease in whom a screening test give a positive result.

Specificity is defined as the ability of a test to correctly identify those free of the disease in the screened population. This is expressed as the proportion of people free of the disease in whom the screening test give a negative result.

These two terms may be further expressed in terms of test results as follows: sensitivity is calculated as the true positives divided by the sum of the true positives and false negatives and may be expressed as a proportion or a percentage; specificity is calculated as the true negatives divided by the sum of the true negatives and the false positives and may be expressed as a proportion or a percentage (Fig. 22.3). In practice, difficulties with these measures arise over defining a positive result from the test as well as distinguishing the true positives from the false positives among those who test positive and the true negatives from the false negatives among those who test negative. A relatively imperfect test of a quantitative, continuously distributed measurement can be artificially given a very high sensitivity by setting the boundary between negative and positive to incorporate a high proportion of those who are eventually found to have the disease in the positive category, but at a substantial cost in terms of low specificity. Conversely the same test can be made to appear highly specific, but will then become insensitive, if the boundary between positive and negative is shifted in the opposite direction.

If the test result is expressed in a quantitative form so that the boundaries between what is defined as positive and negative can be varied at will, it is possible to plot a receiver operating characteristic (ROC) curve (Swets 1979). What is plotted is the sensitivity in the vertical axis and 1-specificity (the false-positive rate) in the horizontal axis (Fig. 22.4). The point on the curve that is chosen as optimal is often that furthest from the 45° diagonal, labeled “Chance” in Fig. 22.4, as this represents

	Disease status	
Test result	Present	Absent
Positive	True positive (*TP*)	False Positive (*FP*)
Negative	False negative (*FN*)	True negative (*TN*)

Then:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \text{ x } 100\%$$

$$\text{Specificity} = \frac{TN}{TN + FP} \text{ x } 100\%$$

$$\text{Predictive value positive} = \frac{TP}{TP + FP} \text{ x } 100\%$$

$$\text{Predictive value negative} = \frac{TN}{TN + FN} \text{ x } 100\%$$

Fig. 22.3 The relationship between sensitivity, specificity, and predictive value positive and negative

a test with no better sensitivity or specificity than could be expected by chance. In Fig. 22.4, this point is between 70% and 80% sensitivity, the corresponding 1-specificity being 25%, i.e., a specificity of 75%. This is about the level expected with cervical cytology and represents rather a poor test. We would hope that even at greater sensitivity, the specificity would be at least 90% to avoid too many health-care costs investigating the false positives. ROC curves are most easily derived for blood tests, but have also been applied to mammography, by varying the extent to which different mammographic abnormalities were regarded as an indication of suspicion of malignancy (Goin and Haberman 1982). Such curves cannot be applied to a test with a dichotomous outcome if the boundaries defining positive and negative are invariant. Further they imply a similar weight to sensitivity and specificity, which, as discussed below, may not be ideal.

The position of the boundaries that are set between what is regarded as disease and non-(or benign) disease can also considerably influence the numerical values placed on sensitivity and specificity. This arises because of uncertainty as to what

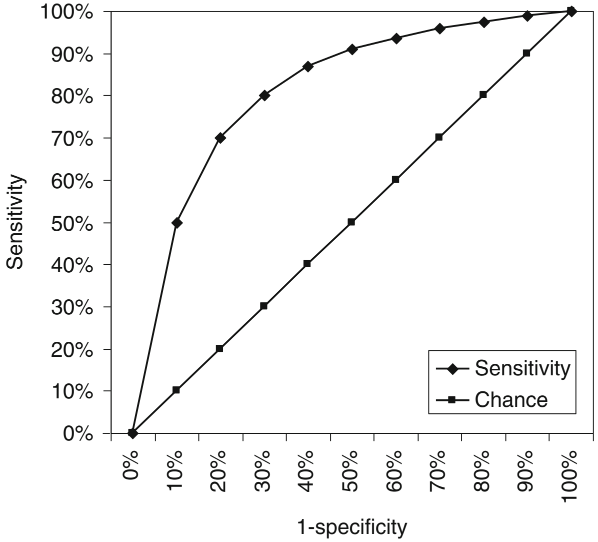


Fig. 22.4 Example of a ROC curve

truly constitutes an abnormality in the context of a screening program. In order to come to such a decision, it is essential that the conditions identified as a result of screening should have a known natural history. However, as has already been pointed out, such knowledge may not be available at the initiation of a screening program and may only be obtained as a result of careful study of findings from screening programs.

Nevertheless, the definition as to what constitutes disease is crucial in order to determine sensitivity and specificity. Most people have a clear idea as to what they regard as disease in terms of that which surfaces in standard medical practice. By definition, screening is conducted on asymptomatic individuals so that many conditions that are identified through screening are likely to be at an early stage and may not have the generally recognized clinical characteristics of relatively advanced disease. This difficulty should theoretically be overcome by having clearly defined definitions of disease. However, for cancer, diagnosis is usually made on histology, and histology only imperfectly characterizes behavior, especially for lesions within the DPCP. For cancer, one hope for the future is that some of the markers for prognosis currently being evaluated such as markers of oncogene expression or other markers of DNA change may serve to identify those precancerous or in situ components of the DPCP that are likely to progress.

A common error in evaluating potential screening tests is to determine the sensitivity by utilizing the experience of the test in relation to people who have clinical disease. A test that may appear to be highly sensitive under these circumstances may later be found to be much less sensitive when its ability to detect the DPCP is evaluated. A similar error is substituting an intermediate marker as the gold standard for calculating sensitivity and specificity. For example, the test characteristics for a thyroid assay may be compared to a more accurate assay, rather than the presence of the condition in the individual tested. In the screening context, therefore, sensitivity and specificity may vary according to whether they are estimated for early disease or preclinical lesions, and sensitivity for both should be determined in active screening programs. To do so for specificity is very much easier than for sensitivity. This is because the diagnostic process put in train by a positive screening test generally fairly rapidly identifies those who have the disease and thus distinguishes the true from the false positives. As under most circumstances, the proportion of those who have the disease in relation to the total population screened is low, a very good approximation to specificity is obtained by calculating the proportion of all those who tested negative of the sum of the test negatives and false positives. Including the unidentified false negatives in the numerator and denominator of this expression will in practice introduce little error.

Sensitivity is however a difficult measure to determine initially in a screening program. The reason for this is that the false negatives are not immediately apparent, as there is no justification to retest all the test negatives just to identify a few false negatives. Only by following the total population who screened negative is it eventually possible to identify those who had the disease at the time the test was administered but were not so identified at the time of screening. This is facilitated if test materials are retained, for example, cervical smears or mammograms originally classed negative can be reassessed for those who are found to have disease at the next scheduled screen or who develop disease during the interval between screens. Such reassessments should preferably be made blind to avoid bias. Such an approach was used in the assessment of the sensitivity of the "reader error" for cervical cancer screening (Boyes et al. 1982) and in the assessment of the sensitivity of mammography in a trial of breast cancer screening (Baines et al. 1988).

When test materials cannot be retained, however, such as in the assessment of the sensitivity of physical examination as a screening test for breast cancer, and for what we have called the "taker error and the biological component" of false negatives in cervical cytology, i.e., disease that was indeed present but was not incorporated in the smear or for some reason did not exfoliate (Boyes et al. 1982), a direct identification of false negatives will not be possible. A usual approach is to assume that disease occurring within a certain period are false negatives, an approach we used in estimating the sensitivity of physical examination of the breasts (Baines et al. 1989). However, a possibly more satisfactory approach is to assess the expected detection rate of disease on screening after repeated screens, assuming that most of the false negatives had by then been identified, and to regard the excess disease above this level at the second screen as a measure of the false negatives at the first screen. As a result of such an approach, it was determined that the taker and biological component of false negatives was approximately equal to the directly measured reader error so that the level of sensitivity for cervical cytology approximated to 78% (Miller 1981).

It is generally accepted that:

- The sensitivity and specificity of the screening test to be used should have been evaluated and their expected values stated.
- There should be an acceptable program of quality control to ensure that the stated levels of sensitivity and specificity are attained and maintained.

Quality control involves issues that concern not only the validity of the screening test but also its safety. There is, for example, the need to ensure that radiation exposure does not drift upward in a mammography screening program. Quality control encumbers the training of those who will actually administer and read screening tests, their supervision, and the introduction of procedures to check actively on the extent to which those positive or negative are misclassified.

Quality may suffer because of overwork and boredom. One of the reasons why a recommendation was made to change the frequency of examination for most women in cervical cytology screening programs in Canada was to avoid repetitive rescreening of normal women, with the flooding of laboratories with unnecessary and unrewarding work (Task Force 1976). The Task Force described the mechanisms for ensuring appropriate quality control. It is relevant that these requirements had to be reemphasized more than a decade later (Miller et al. 1991b).

That such issues are not simple was underlined by consideration of observer variation in mammography reading (Boyd et al. 1982). Relevant to all screening programs is not only the accuracy with which abnormalities are identified but, if identified, the extent to which appropriate recommendations are made on their management. Our experience suggested that including a category of "probably benign" in a screening mammography report increases the extent of observer variation. Readers differ substantially in the extent they use this category, the extent to which they recommend special observation of individuals placed in this category, and the extent to which they recommend biopsy. Dual reading helps to increase specificity, without much, if any, loss of sensitivity. This permits the simplification of recommendations into two groups, "suspicious of malignancy" and "satisfactory (normal)" examination, and results in far greater consistency. Further, it is compatible with the appropriate separation of findings from screening tests into the probably abnormal (test positive) and probably normal (test negative) dichotomy. The probably abnormal group is subjected to diagnostic tests in the normal way. This approach to use of screening mammography was accepted with difficulty in North America due to an initial tendency for most radiologists to regard mammography as a diagnostic rather than a screening test. This resulted in greater use of biopsy as a diagnostic test in North America than that reported from Europe (McLelland and Pisano 1992), where more use was made of diagnostic mammography subsequent to screening mammography (often called by European radiologists "complete" mammography) with a consequent reduction in biopsies and a much lower benign to malignant ratio.

Most commentators in the past, when considering the relative weight to be placed on sensitivity and specificity, tended to encourage high sensitivity at the cost of relatively low specificity, as it was felt important to attempt to avoid missing individuals who truly had disease. One vigorous exponent of this view for breast cancer screening was Moskowitz, who coined the term "aggressive screening," as only by such an approach did he feel that the "minimal" breast cancers with an excellent prognosis would be identified (Moskowitz et al. 1976). However, there continues to be little evidence that such cancers are really responsible for the mortality reduction following breast cancer screening. Rather, there is much evidence that the early diagnosis of more advanced disease results in the benefit (Miller 1987, 1994; Miller et al. 2000b, 2002). A disadvantage of aggressive screening was a high benign to malignant ratio and low specificity of the screen. Although the objective of screening is to identify disease in the DPCP before it gets to the stage of escaping from curability, if a test is made so sensitive that it picks up lesions that would never have progressed in that individuals lifetime, there will be substantial additional costs for diagnosis and treatment (this is one consequence of the "overdiagnosis" bias, which is more fully discussed in relation to survival of cases following screen detection in a later section of this chapter). There is no point in identifying through screening disease which would never have presented clinically and little point (other than less radical therapy) in identifying early disease that would have been cured anyway if it had presented clinically. Similarly, identifying disease that results in death, even following screen identification and subsequent treatment, only results in greater observation time and no benefit to the screenee. It is only disease that results in death in the absence of screening, but which is cured following treatment after screen detection, from which the real benefit of a screening program derives. Hence, if high "sensitivity" is largely based on finding more good prognosis disease, but results in lowering specificity, the program will incur much greater costs without corresponding benefit.

The process measure, as distinct from a measure of validity, that most clearly expresses this difficulty is the predictive value of a positive screen. This is defined as the proportion of those who test positive who truly have the disease (Fig. 22.3). This measure is influenced not only by the sensitivity and specificity of the test, especially the latter, but by the prevalence of disease in the population, whereas sensitivity and specificity are invariant with regard to disease prevalence. If tests are administered under circumstances that incur a low predictive value positive, then not only may costs be high in terms of correctly identifying those who are falsely positive but also the potential hazard may be high, as an individual classified as positive falsely derives no benefit and potentially a substantial risk from the associated diagnostic procedures. A test with a low positive predictive value rapidly enters into disrepute. The predictive value positive will be very much higher when people at high risk for the disease are screened, as distinct from the general population. This is the main reason why in screening programs, those eligible are defined by age.

This is sometimes misunderstood by those responsible for organizing screening, especially if their experience to date has been largely related to the treatment of the disease for which it is now intended to screen. As an example, because of the different age pyramid in most developing countries, those who treat breast cancer in such countries often believe that the cancer they see is different from technically advanced countries, as it occurs a decade earlier. The error has been to look at the age distribution of cases and not relate that to the age distribution of the population from which the cases are drawn. So that often, such clinicians will demand screening to begin in women in their thirties, rather than in their forties or fifties as in developed countries, as a much higher proportion of the cases they see are young than in developed countries. What they have failed to understand is that the age-specific incidence of breast cancer in women in their thirties is no higher than in technically advanced countries and is often lower. So that, to take an example from one country in the Eastern Mediterranean region of the World Health Organization, if they were to attempt to screen women at, say, age 35–39, they would on average have to examine over 3,500 women to detect one case of breast cancer, whereas the burden would be much less if they were to screen older women. This relationship between prevalence and positive predictive value is causing concern as we begin to contemplate the requirement for continued screening of women vaccinated to prevent cancer of the cervix, given that the available vaccines do not provide protection against all high-risk HPV types (Howlett et al. 2009). Although screening will be needed in such women, the positive predictive value will be much lower than in unvaccinated, even normal risk women.

To complete discussion of process measures, the predictive value negative should be defined. This is the proportion of those who test negative who are truly free of the disease. This measure, like sensitivity, is dependent on identifying the false negatives and therefore is rarely determined while being of little operational value. In practice, however, it is usually high. However, it has, in theory, become important in relation to cervical screening, where one of the major benefits of using primary HPV screening is likely to be its higher negative predictive value than cytology, thus permitting longer intervals between repeat screens (Dillner et al. 2008).

As Day (1985) has pointed out, because of the difficulty in identifying false negatives, and because of the overdiagnosis bias, the usual approach to defining sensitivity is not ideal nor particularly biologically meaningful. He suggested an alternative measure of sensitivity which can be derived if the expected incidence of disease in the absence of screening can be determined, ideally from the control group in a randomized trial but sometimes in population-based programs from historical data or data from comparable unscreened populations. The method basically computes the extent a program is successful in reducing the expected incidence of disease in the absence of screening. The lower the proportion of expected incidence occurring after screening, the greater the sensitivity. The use of this approach, and other measures relevant to sensitivity, such as test, episode, and program sensitivity, is fully discussed by Hakama et al. (2007).

22.2.2 The Acceptability of the Test

One of the desirable attributes of a good screening test is that it should be acceptable to the population to which screening is offered and acceptable to those who will administer the test. In general, cervical cytology screening programs have found acceptance with women and their physicians. Although those who tend to be at highest risk of the disease do not comply as well as those with lower risk, this is not strictly related to acceptability of the test, but rather the fact that such individuals tend to have so many pressing health and social problems, that taking action to reduce an uncertain risk in the future does not have priority with them. Nevertheless, this results in lower effectiveness of programs than would be the case if all women were to be included. This lack of acceptance is largely related to lower socioeconomic status, accompanied by high parity, and often multiple sexual partners.

Breast cancer screening has encountered different problems over acceptability, though this varies substantially in different countries, ranging from the 90% acceptance with screening invitations in the Swedish Two-County Trial (Tabar et al. 1985) to the difficulties with both physician and women compliance in the early stages of mammography utilization for breast screening in the United States (Howard 1987). In Europe, a median uptake of 74% has been reported for mammography offered in the population (European Society for Mastology 1993). Perhaps fortunately, those who tend to comply with invitations to attend breast screening programs tend to be those at higher risk. An exception relates to the misperception of many older women that they are not at risk for the disease, whereas the older they become, the greater is their risk.

Another example is screening for colorectal cancer because of the inevitable distaste of individuals for a procedure that involves manipulation of feces. In a number of pilot programs, therefore, the return rates for hemoccult slides have been low, though they have been better in well-organized studies (Chamberlain and Miller 1988), and achieved approximately 75% in the Minnesota Colon Cancer Screening Trial (Mandel et al. 1999).

A screening test, therefore, has to be acceptable to the population in its widest sense. The test should be simple and as far as possible easily administered. It should involve procedures that are not unacceptable, and its use should not have unpleasant or potentially hazardous implications. There are also economic advantages in a test being administered or read by allied health professionals, such as use of technologists in screening cervical cytology slides (Anderson 1985) or the use of nurses to perform breast examinations (Bassett 1985; Miller et al. 1991a).

22.3 The Ethics of Screening

In general medical practice, the special nature of the relationship between a patient and her or his physician has dictated the need to build up a core of ethical principles that govern this relationship. Further, it is generally accepted that additional issues arise when a patient becomes the subject of a research investigation that is superimposed upon his or her search for and receipt of appropriate medical care. It was not initially appreciated, however, that screening opened up a completely new spectrum of issues, possibly requiring more restrictive boundaries of ethical behavior than those applied in usual medical care. For example, when a patient goes to see a physician for relief of a symptom or treatment of an established condition, the physician is required to exercise his or her skills only to the extent that knowledge is currently available, while doing what is possible with available expertise and appropriate assistance to help the patient. Treatment may be offered without any implied guarantee that it is necessarily efficacious or will do more than just temporarily relieve the symptoms of which the patient complains. Thus, the physician promises to do his or her best for the patient; there is no implied promise that the patient will be cured.

In screening, however, those who are approached to participate are not patients, and most of them do not become patients. The screener believes that as a result of screening, the health of the community will be better. He or she does not necessarily intend to imply that the condition of every individual will be better. However, screening is often promoted as if it implies a benefit to everyone who is screened. In fact, in some circumstances, individuals included in a screening program may be placed at a disadvantage, as discussed above. Furthermore, the harm from a screening test is not only related to the risk of being false positive or negative. Those that are screened may also incur psychological consequences, sometimes merely from being labeled as being at risk of disease. At the very least, therefore, those planning to introduce a screening program should be in a position to guarantee overall benefit to the community and a minimum of risk that certain individuals may be disadvantaged by the program. It was the inability to guarantee overall benefit and lack of disadvantage for those screened that led to the proscription of mammography in women under the age of 50 in the Breast Cancer Detection Demonstration Projects in the United States (Beahrs et al. 1979).

A second ethical issue, which is directed more to the obligations for appropriate care in the community than toward individuals, concerns how limited resources are equitably distributed across the whole community to obtain maximum benefit. Under certain circumstances, the offer of screening could diminish the total level of health in a community. This may be a particular problem for developing countries by diverting resources intended for routine health care into screening. Thus, resources diverted to a screening project, which might be regarded as prestigious, especially if involving high technology, could lower the resources available for other more pressing but also more mundane health problems. Although several screening programs have been proposed for developing countries, there is a particular need for caution and care in order to ensure that they do not overbalance the health-care system in the area in which they are introduced.

A final ethical dilemma for screening programs is how to implement informed consent. Information about risks and benefits of tests and treatments is expected to be provided in usual clinical practice. For screening, providing information about the test alone is not sufficient. Information about the consequences of the test, the diagnostic assessment process, and the diseases to be detected and their treatments should also be presented, if a truly informed decision is to be made. Presenting such a large amount of information is obviously difficult, particularly in a primary care setting where several screening tests may be done at the same time. Furthermore, presenting such information becomes even more cumbersome when the evidence base for a screening test is limited, such as in the case with prostate screening with the prostate-specific antigen (PSA) test.

Those in charge of screening programs, therefore, carry an ethical responsibility at least as great as that for medical practice in that approaches to participate are made to ostensibly healthy people. Indeed, the burden of proof for efficacy of the procedures and the necessity to avoid harm are greater than may be required for diagnostic or therapeutic procedures carried out when a patient presents with symptoms to a physician. In screening, the physician or public health worker initiates the process, and he or she bears the onus of responsibility to be certain that benefit will follow.

22.4 The Population to be Included in Screening Programs

For a screening program to be successful, the population to be included should be one in which it is known that the disease has a high prevalence. This will not only encourage a high predictive value for a positive test; it will tend to promote higher quality of performance and assessment of results of screening tests and will result in lower costs per case detected. Thus, in all screening programs, it is desirable to attempt to include only those who are at risk of the disease and to concentrate particularly on those who are at high risk of the disease. This approach was recognized by the Canadian Task Force on Cervical Cytology Screening Programs (Task Force 1976) who carefully defined those whom it believed were at such low risk for the disease that they need not be included in cervical cytology screening programs, thus defining the remaining "at-risk" population on whom major efforts should be concentrated to bring them into screening. In the case of cardiovascular disease, family history and factors such as smoking have been used by some to define populations to be screened.

For other diseases, however, the known risk factors, apart from age, may not suffice to adequately distinguish between those who should be considered for inclusion in screening programs compared to those who should not. For breast screening, for example, although some discrimination using risk factors has been achieved (Schechter et al. 1986), this has not been sufficient to justify selection on this basis alone. However, age is an important predictor of risk, and for breast cancer in technically advanced countries, all women in the appropriate age group can be regarded as at high risk. Thus, for breast cancer currently, it seems unlikely that any program could justify routine screening of women under the age of 40, while several trials only found a delayed effect of screening women in this age group, and our trial in Canada, the only so far specifically designed to evaluate breast screening at age 40–49, could not find evidence that screening women at these ages with mammography is effective (Miller et al. 2002). The other trial designed to evaluate screening at these ages had an innovative design, women entering their forties were randomized 1:2 to annual mammography screening for 7 years compared to no screening, followed without screening until they reached the age of 50, and then all were invited to participate in the UK screening program of 3-yearly mammography also found no significant benefit from screening at these ages (Moss et al. 2006). As a result, the organized programs in most European countries and in Canada do not attempt to recruit women under the age of 50 into screening.

One possible approach to concentrating on the relevant segment of the population for screening might be to administer a prescreening test, especially for a marker for a factor necessary in the causation of the disease. The test for human papilloma virus infection could be used as a prescreen for cancer of the cervix (Miller et al. 2000a), as the test does not identify the presence of disease – other tests such as cytology of colposcopy have to be used in those found positive to the test. A difficulty here is the high proportion of infections that are self-limiting without the development of high-grade cervical intraepithelial neoplasia. This means that the test is too non-specific if used among women under the age of 30. However, in older women, if found to be negative to a test for oncogenic HPV strains, they can be deemed at low risk for the disease and screened far less frequently than women who are HPV positive (Dillner et al. 2008).

The development of genetic susceptibility testing opens up the possibility in the future of prescreening for a range of diseases. Unfortunately, the tests that indicate high risks so far only identify individuals in the population responsible for a low proportion of disease. It seems possible that it will be necessary to screen for a combination of genetic polymorphisms each responsible for relatively low relative risks in the population, but with high attributable risk, in order to adequately discriminate at risk from low-risk individuals. This will increase the complexity, and costs, of the process considerably.

Nevertheless, in programs that attempt to select for screening on the basis of risk factors, there will almost invariably be cases occurring in the unscreened group. A consequence of such programs, however, will be reduced numbers of false positives (in absolute terms) which with the increased prevalence of the disease will result in a higher positive predictive value of the screen. Hakama (1984) coined the terms program sensitivity and program specificity which help in understanding the effects of screening concentrating on “high-risk” subjects. The more a program concentrates on high-risk groups, the lower the program sensitivity, as more and more cases will occur in unscreened people. Conversely, however, the program specificity will increase because of the increase in healthy people unscreened with a reduction in the costs of screening. However, because the reduction in program sensitivity will result in a reduction in the overall effectiveness of the program, the overall result of such an approach could be unacceptable.

One other approach to using risk factors is to help determine the optimal periodicity of rescreening. Those judged at high risk could be screened more frequently, as in the example of HPV tests for cervix screening given above. Once again, however, much of the necessary research is incomplete, and we do not know how appropriate such an approach may be. It will probably be necessary to calculate the marginal cost-effectiveness of extending screening from high- to low-risk groups (i.e., the additional cost for such an extension of screening related to the increase in effectiveness of the screening) in order to make the necessary policy decisions.

22.5 Diagnosis and Treatment of the Discovered Lesions

As a screening test is not diagnostic, inevitably the success of the program will ultimately depend upon the extent those identified as having a positive test accept the procedures offered to them for further evaluation and the effectiveness of the therapy offered.

A number of difficulties may arise. For example, in the initial phases of many breast cancer screening programs, it was necessary to demonstrate to the general community of medical practitioners that the abnormalities identified were indeed of importance and that they required care and expertise to biopsy. Indeed in the absence of skills in diagnosis and management, there can be unnecessary biopsy (potentially reducible by the use of diagnostic mammography and fine needle aspiration biopsies) as well as failure to excise the lesion when biopsy is performed. This is part of the spectrum of problems that arise over the fact that lesions may be identified in screening programs whose biological features, natural history, and other characteristics may be in doubt. The screening participants may require special education programs so that they understand the diagnostic process to reduce as far as possible one of the major adverse consequences of screening, the anxiety accompanying the identification of an abnormality, as well as ensuring that they comply with the recommendations for management. There may even be major disagreements over the histological interpretation of the excised lesions, with uncertainties over the borderline between benign and malignant. Thus, the public and the professionals at all levels in a screening program may require education and/or retraining dependent on their responsibilities. One mechanism to reduce difficulties in the professional area that should be encouraged is the provision of special diagnostic and treatment centers where the necessary expertise in diagnosis and management can be concentrated and where the necessary facilities are available (Miller and Tsechkovski 1987). Such diagnosis centers could be regionally based, serving a number of screening centers.

22.6 Evaluation of the Efficacy of Screening

A number of issues have to be noted when evaluating the efficacy of screening. Almost invariably individuals with disease identified as a result of screening will have a longer survival time than those diagnosed in the normal way. Four biases associated with screening explain this (see Box 22.1). The first is “lead time,” defined as the interval between the time of detection by screening and the time at which

Box 22.1. The biases associated with survival

Lead time: the period by which screening advances the diagnosis of the disease.

Length bias: less rapidly progressive cases are likely to be detected by screening.

Selection bias: volunteers for screening are likely to have a better outcome from their disease than those who decline screening.

Overdiagnosis bias: lesions identified by screening which would not have progressed to clinical disease in the absence of screening.

the disease would have been diagnosed in the absence of screening (depicted in Fig. 22.1). In other words, it is the period by which screening advances the diagnosis of the disease. For example, if, as a result of screening, the average point of diagnosis is advanced by 1 year, then inevitably cases diagnosed by screening will survive 1 year longer even if there is no long-term benefit. It is important to recognize that the lead time for different cases will vary, depending in part on the timing of the screening test in relation to the duration of the DPCP in that case, as well as the rapidity of progression of the DPCP in that individual. Thus, there will be a distribution of lead times (Morrison 1985). The lead time for fatal cases will be fairly short, but in one study, some fatal cases have been identified as having a lead time of one or more years following mammography screening (Miller et al. 1992).

The determination of lead time is complex, but models have been developed that do so providing there are control data that permit comparison of screen detection with that expected (Walter and Day 1983).

Differential lead time can be an important factor in comparing the outcome among cases detected by different screening modalities, making it almost impossible to make a comparison based on survival, unless it is possible to estimate and correct for differential lead time (Walter and Stitt 1987).

The second bias that accounts for improved survival of screen-detected cases is "length-biased sampling," or more simply, length bias. This relates to the fact that individuals who have rapidly progressive disease will tend to develop symptoms that cause them to consult physicians directly. Thus, only less rapidly progressive cases are likely to remain to be detected by screening. Yet the former have a poorer and the latter a better prognosis, hence the improved survival of screen-detected cases, over and above lead time. A representation of this bias is in Fig. 22.5. It is obvious that if the screening test were to be administered just before clinical symptoms occur, nearly all cases would be detected, but by them, some of them would probably have a poorer prognosis. So it is important to try and detect cases earlier in the DPCP, but if one is successful, many of the rapidly progressive cases will be missed. What is more, what is depicted is just at one point in time, rapidly progressive cases that were present at the time the slowly progressive cases become detectable, would have

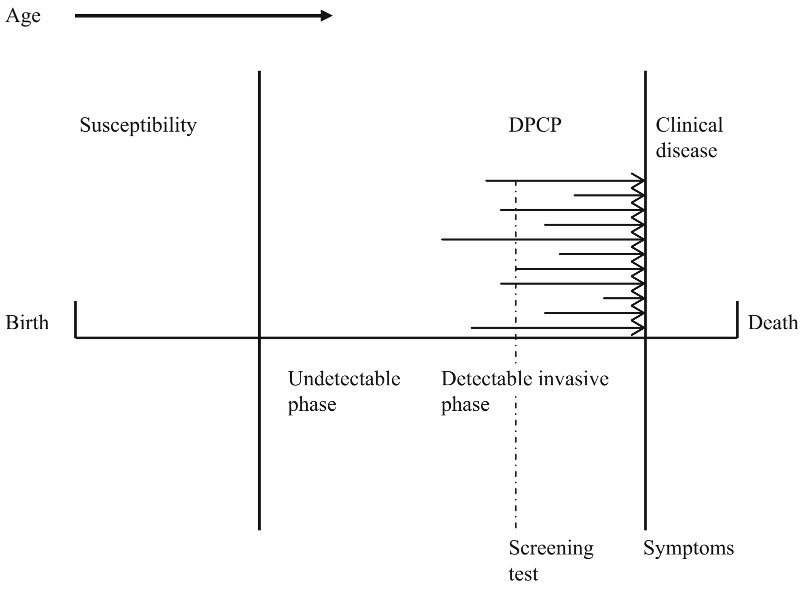


Fig. 22.5 The impact of a screening test when the detectable preclinical phase varies in duration – length bias

presented with symptoms before the screening test was administered. This bias is most obvious at the initiation of a screening program, at the first or prevalent screen. However, length bias will also affect the type of cases detected at rescreening, with the more rapidly progressive cancers diagnosed in the intervals between screens. Hence, in evaluating the total impact of programs, the interval cases must be identified and taken into consideration as well as the screen-detected cases.

The third bias which can artifactually improve survival is selection bias, i.e., participants in screening are likely to have a better outcome from their disease than those who decline screening. This is because those who enter screening programs are volunteers and almost invariably more health conscious than those who decline to enter. This means that they are likely, even in the absence of screening, to have a better outcome from their disease than the overall rates in the general population.

The fourth bias is overdiagnosis bias. Simply it means that some lesions identified and counted as disease would not have progressed to present clinically in those individuals during their lifetimes in the absence of screening. It is, in practice, an extreme example of length bias. It is difficult to obtain absolute confirmation of the existence of this bias, though it seems likely that it is at least in part an explanation for the substantial excess of cancers detected by PSA screening for prostate cancer (Draisma et al. 2003), and it has also been strongly suspected for breast cancer (Miller et al. 2000b, 2002).

The only design that effectively eliminates the effect of all these biases is the randomized controlled trial (Prorok et al. 1984), but only if mortality from the disease (i.e., deaths related to the person-years of observation) is used as the endpoint, rather than survival. Survival could be used in a randomized controlled screening trial only under special circumstances. These are that there is good evidence because of the equivalence in cumulative numbers of cases during the relevant period of observation that there is no overdiagnosis bias and providing that the start of the period of observation of the cases is taken as the date of randomization, as that will eliminate differential lead time. This is the approach that was planned to be used in a study of breast self-examination in Russia, where it will not be possible to follow all entrants to determine their alive and dead status at the end of the trial (Semiglazov et al. 1993). Length bias and selection bias are not issues, the latter having been equally distributed by the randomization and the former by having started at the same point in time and by including all cases that occur during follow-up in the evaluation. However, to date, it has not been possible to perform a definitive analysis of the results of this trial, as the collapse of the Soviet Union resulted in too much disruption to permit the collection of the data deemed necessary to adjust for the cluster (as distinct from individual) randomization performed in the trial.

Outside a randomized trial, if the screening test detects a precursor, reduction in incidence of the clinically detected disease can be expected and evaluated. This effect has been well demonstrated in the Nordic countries in relation to screening for cancer of the cervix (Hakama 1982). It is also anticipated from endoscopy screening for colorectal cancer, with the detection and removal of adenomas, and was demonstrated in the randomized Minnesota trial of colorectal screening using the fecal occult blood test with colonoscopy extensively used for diagnosis (Mandel et al. 2000). If the screening test does not detect a precursor, or even if it does but the main yield is invasive cancer, then incidence can be expected to increase initially following the introduction of screening and remain elevated while screening continues, though there may be some reduction toward the baseline after continued screening, if the application of the test results in most at-risk subjects being included, and the subsequent screening tests are largely used for rescreening. This seems to have happened with the major rise and then fall, but not to the original levels, of prostate cancer incidence following the opportunistic introduction of the PSA test in North America. Under such circumstances, when further reduction in incidence cannot be anticipated, and improvement in survival cannot be relied upon because of the biases already discussed, the only valid outcome for assessment of results of a screening program is mortality from the disease in the total population offered screening in comparison with the mortality that would be expected in the same population if screening had not been offered.

As already emphasized, the design of choice for evaluation of changes in mortality is the randomized controlled trial. This can either be an efficacy trial or an effectiveness trial. Efficacy trials are based on randomization of the screening test, which answers the biologically relevant question as to whether mortality is reduced in those screened. An effectiveness trial is based on the randomization of invitations to attend for screening and more nearly replicates the circumstances that may eventually pertain in practice in a population. Both those who accept the invitation as well as those who refuse will have to be included in the assessment of outcome. Thus, it tests the impact of introducing screening in a population. Some trials of this type involve randomization by cluster. However, cluster randomization can lead to difficulties in determining whether the trial results are valid, especially if it cannot be confirmed that the randomized groups were balanced, or if there is evidence that they were not. Such problems led to difficulties in determining the validity of some cluster randomized trials of breast screening (IARC 2002).

If for some reason randomization is believed inappropriate, a second-best method is the quasi-experimental study in which screening is offered in some areas, and unscreened areas as comparable as possible are used for comparison purposes. However, this design is not a cheap and easy way out but demands the same methodological accuracy as required for randomized trials. Further, in view of the substantially larger populations that may have to be studied than in randomized trials, it may prove to be more expensive than the preferred design. Critically, difficulties in analysis may ensue if the baseline mortality in the comparison areas differs (UK Trial of Early Detection of Breast Cancer Group 1988).

Nevertheless, ethical issues may preclude the utilization of randomized trials, particularly for programs that were introduced before the necessity of utilizing trials as far as possible for evaluation was appreciated. This has been the case for screening for cancer of the cervix, for example. One approach under these circumstances is to compare the mortality in defined populations before and after the introduction of screening programs, preferably with data available on the trends in acceptance of screening so that changes in mortality can be correlated with the mortality trends. Such a correlation study will be strengthened if other data that could be related to changes in the outcome variable are entered into a multivariate analysis (Miller et al. 1976).

A case-control study of screening is another approach that can be used to evaluate programs that were introduced sufficiently long before the study that an effect can be expected to have occurred. Case-control studies depend on comparing the screen histories of the cases with the histories of comparable controls drawn from the population from which the cases arose. Individuals with early stage disease if sampled would be eligible as a control, providing the date of diagnosis was not earlier than that of the case, as diagnosis of disease truncates the screening history. However, a bias would arise if advanced disease is compared only with early stage disease, as the latter is likely to be screen-detected, though this is just a function of the screening process, not its efficacy (Weiss 1983). Cases have to reflect the endpoints used to evaluate screening, i.e., those that would be expected to be reduced by screening. Thus, cases are often deaths from the disease or advanced disease as a surrogate for deaths or, if a precursor of the disease is detected through screening, incident cases in the population. If incident cases are screen-detected, the controls should be drawn from those screened in the same program; if the cases are not screen-detected, the controls should be population-based (Sasco et al. 1986).

One difficulty with case-control studies of screening is that they may be affected by selection bias as the health conscious may select themselves for screening. This may be difficult to correct in the analysis, though such a correction should be attempted if the relevant data on risk factors for the disease (confounders) are available. This was possible in a case-control study of breast self-examination nested within the Canadian National Breast Screening Study, in which data on risk factors for breast cancer were collected from all on enrollment (Harvey et al. 1997). Such a bias may not be a problem, however, in other circumstances, if it can be demonstrated that the incidence of cancer in those who declined the invitation to the screening program is similar to that expected in an unscreened population.

However, even if data are available on risk factors for disease, adjusting for them may not result in avoiding the effect of selection bias. For breast cancer, for example, experience in studies in Sweden and the UK, where case-control studies were performed within trials, shows that although those who refused invitations for screening had similar breast cancer incidence to the unscreened controls in the trial, their breast cancer mortality experience was worse than that of those controls. This meant that the estimate of the effect of screening in such case-control studies was greater than could have been expected in the total population (Miller et al. 1990; Moss 1991).

In addition to assessing effectiveness of screening, case-control studies may also be of use to assess other aspects of screening programs. For example, a method has been proposed for estimation of the natural history of preclinical disease from screening data based on case-control methodology (Brookmeyer et al. 1986).

The cohort study design may also provide an estimate of the effect of screening, an approach in which the mortality from the cancer of interest in an individually identified and followed screened group (the cohort) is compared to the mortality experience in a control population, often derived from the general population. This approach has been used to evaluate the mortality experience in the US Breast Cancer Detection Demonstration Project (Morrison et al. 1988) and in a cohort of women in Finland included in a breast self-examination program (Gastrin et al. 1994). In these studies, it has to be recognized that those recruited into a screening program are initially free of the disease of interest so that it is not appropriate to apply population mortality rates for the disease to the person-years experience of the study cohort. Rather, as is required in estimating the sample size required for a controlled trial of screening, it is first necessary to determine the expected incidence of the cases of interest, then apply to that expectation the expected case-fatality rate from the disease to derive the expectation for the deaths (Moss et al. 1987). In practice, a cohort study of screening suffers from the same problem of selection bias as for case-control studies, so the results have to be interpreted with caution.

Indirect indicators of effectiveness are often desired in evaluating screening programs, especially one that would predict subsequent mortality. Compliance with screening, and rate of screen detection, as well as the ratio of prevalence and incidence can be indicators of potentially effective screens (Day et al. 1989). The cumulative prevalence (not the percentage distribution) of advanced disease is one such measure (Prorok et al. 1984). For example, reduction in advanced disease predicted subsequent breast cancer mortality reduction in a trial of mammography screening versus no screening in Sweden (Tabar et al. 1989). However, case detection frequency, numbers of small tumors, and stage shift in percentages of the total should not be used as indicators of effectiveness as they potentially reflect all four screening biases.

In evaluating whether screening programs are effective in a population, different methods have generally to be used. They are considered in the Sect. 22.11, later in this chapter.

22.7 Organized Screening Programs

There are a number of features of effective screening programs that are largely related to good organization. Indeed, there is good evidence, at least for cancer of the cervix, that unorganized or opportunistic screening programs, which depend on the willingness of individuals to volunteer for screening, and the extent to which their physicians offer screening, often to low-risk women, are far less successful (Hakama et al. 1985).

Hakama et al. (1985) defined certain essential elements of organized programs. These are:

- The target population has been identified.
- The individual women are identifiable.
- Measures are available to guarantee high coverage and attendance such as a personal letter of invitation.
- There are adequate field facilities for performing the screening tests.
- There is an organized quality control program on performing and reading the tests.
- Adequate facilities exist for diagnosis and for appropriate treatment of confirmed abnormalities.
- There is a carefully designed and agreed referral system, an agreed link between the participant, the screening center, and the clinical facility for diagnosis of an abnormal screening test, for management of any abnormalities found and for providing information about normal screening tests.
- Evaluation and monitoring of the total program is organized in terms of incidence and mortality rates among those attending, among those not attending, at the level of the total target population.
- Quality control of the epidemiological data should be established.

Although these elements are present in many European cancer screening programs, especially in the Nordic countries, and contribute greatly to their success, several elements are missing from programs elsewhere, especially those largely based on the private medical care system in North America. In Canada, there are opportunities for introducing some of them, such as the first three, and these were recommended by the two Canadian Task Forces on cervical cancer screening

(Task Force 1976, 1982). Unfortunately, only three of the provincial health-care authorities (Ontario, Manitoba, and British Columbia (the latter having accepted from the beginning the need for centralized laboratory services)) have taken the initiative to establish such programs. All provinces in Canada that introduced breast screening programs, however, accepted from the outset the necessity for them to be organized (Workshop Group 1989), thus attempting to replicate the organization of breast cancer screening in some of the Nordic countries, the Netherlands, and in the United Kingdom.

There has been some debate as to whether organized as distinct from opportunistic screening is most efficacious in reducing the incidence of cancer of the cervix. Nieminen et al. (1999) produced evidence that in Finland, the organized program is far more effective than opportunistic screening.

22.8 Health-Related Quality of Life and Screening

An important evaluation measure for screening is the extent overall quality of life is improved or impaired by screening compared to usual care. Decision making for health-care policy is only possible if information is available on quality of life as well as health costs of screened and unscreened participants as well as mortality reduction from screening. For example, it requires an “optimistic” estimate of screening effectiveness to derive an overall benefit from screening for prostate cancer (Krahn et al. 1994). Issues concerning health-related quality of life (HRQL) may well vary with different cultural value systems and different health-care systems.

Because of lead time, HRQL events will tend to occur earlier in life than similar events associated with usual care. Given that the adverse quality of life associated with false-positive screening tests, and those associated with treatment will tend to occur relatively early, it could be easy to convince oneself (as it has convinced some commentators for prostate cancer screening already) that the HRQL issues are overwhelming and that screening should not be conducted. It will require prolonged follow-up, probably more than 10 years, for the detriments associated with advanced disease late in life that may be prevented from occurring in the screened group to appear in the non-screened group (Miller et al. 2001).

If the outcome of screening were to be a major benefit in terms of mortality reduction, the issues related to HRQL would be overwhelmed. It is only if the outcome is a moderate to small mortality reduction that these issues become critical, and paradoxically then it would be necessary for them to have been measured with as much precision as was possible during screening, as, particularly for HRQL, the decrements could not be measured retrospectively with precision. For this reason, in screening trials where adverse HRQL can be anticipated, it is important for such events to be identified and quantified.

22.9 Economics of Screening

Space does not permit a detailed evaluation of the various principles that have to be considered in assessing the economics of screening. In brief, it is necessary to determine the costs of the test and the subsequent diagnostic tests. Also should be included are the costs associated with any hazard of the test as well as the costs of overtreatment. Balancing these costs may be reduced costs of therapy of the primary condition, reduced costs associated with less expenditure on the treatment of advanced disease, and the economic value of the additional years of life gained. This can become quite complex when the value of treatment of disease in years of life gained, transfers such as pensions, and economic productivity are considered. The latter is often disputed, if not regarded with some distaste, so that often what is computed is the cost per year of life saved. Critical may be the marginal costs of additional tests in relation to the benefit, especially when considerations of the frequency of rescreening arise.

Part of the difficulty in economic assessment is that costs are often incurred early, while benefits flow later, so that for proper comparisons of such costs, they have to be discounted to the present day. Additional complexity ensues if attempts are made to assess quality of life in economic terms, while the calculations rarely attempt an economic assessment of the fact that if a death is prevented by screening, the relevant individual will inevitably die of some other condition and that death could be more costly.

It is likely that economic assessments will increasingly guide policy decisions in the future, so it behooves those interested in evaluation of screening to collect the necessary data. Although some economic assessments have suggested that cost-effective programs are achievable, for example, programs of breast cancer screening using single-view mammography in Sweden (Jonsson et al. 1988), others have suggested that programs may not be cost-effective, for example, breast cancer screening programs for younger women in the US (Eddy et al. 1988). Economic analysis is particularly important for making decisions *within* screening programs, for example around screening intervals or method of follow-up. Economic analyses have also facilitated the planning of national breast screening programs (IARC 2002).

22.10 Genetic Susceptibility Testing

The completion of mapping of the human genome holds great promise for disease control. This presages the advent of genetic susceptibility testing. However, the availability of a range of markers for disease susceptibility will lead to increasing controversies about the use of screening tests. While the general principles of screening outlined above will still apply, they will need to be modified. Genetic screening will identify individuals at risk of disease, rather than those with precursors or early stage disease. Although this could lead to focused application of screening tests on those at higher risk, ideally, primary prevention strategies will be available. Nevertheless, for many conditions, the preventive strategies will be the application of other screening tests implying that diagnostic assessment and treatment strategies will still be required.

22.11 Surveillance

Surveillance of a program is performed to assess its performance and to ensure that it is being as effective as possible. This requires adopting the principles of program evaluation.

Program evaluation is "the systematic assessment of the operation and/or outcomes of a program or policy compared to a set of explicit or implicit standards, as a means of contributing to the improvement of the program or policy" (Weiss 1998). Continuous evaluation of processes and outcomes of a screening program is an essential tool for assessing its organizational progress and enhancing its effectiveness.

Monitoring is the process of ongoing evaluation to determine whether a program is achieving its intermediate objectives. Monitoring uses "process measures," that are designed to indicate whether the program is on course to achieve its objectives. Of themselves, these process measures are not indicators of the success of the program. Rather they indicate whether or not the program is on the route to ultimate success, as unless they are achieved, the program is unlikely to be successful. These process measures cover a number of different aspects of the program. If the targets, as outlined below, are not met, it should be obvious which aspects of the program require urgent attention by the program managers to rectify the situation. Monitoring should be an ongoing activity, carried out as an integral part of the program, by the program's own staff.

Evaluation is a process whereby the success of the program in reaching the targets set for the intermediate outcome and outcome measures is carefully reviewed and analyzed. When carried out during the early phases of the program, it will indicate whether the program is likely to be successful. When carried out at certain defined time points after the initiation of the program, it will indicate whether the program has been successful, and, whether in the light of the degree of that success, it should be continued. Evaluation is an intermittent activity. It can be carried out by the program staff, but it can also be performed by an external consultant to the program, providing the information system is fully functional and successful in obtaining the data required.

Surveillance and evaluation are essential for all programs to ensure that the resources used achieve the benefit expected. Process measures such as the numbers of screens performed, the numbers of positive tests reported, the number (and proportion) of those screened referred for diagnosis and therapy, the numbers of cases of disease diagnosed, and the numbers of precursor and benign lesions detected should be derivable providing the data are collected from the screening centers and the treating institutions and collated. Such data must be analyzed by age to confirm that those in the target age group are being screened and receive appropriate subsequent management. However, such data cannot evaluate the effectiveness in terms of the likely prevention of occurrence of disease or of deaths from the disease, unless the data can be related to that derived from the total population on an ongoing basis, which requires linkage to a preexisting disease register or vital statistics system, or to a register of cases of the relevant disease established for this purpose.

Depending on the endpoint that should be affected by screening, e.g., deaths from the cancer of interest for cancer screening programs, the simplest form of surveillance and evaluation that will provide measures of the effectiveness of the program in the population is to be able to demonstrate a change in the slope of the trend in mortality from the disease in the population. More detailed evaluation requires the identification of all who develop the disease and die from it in the target population and documentation of their screening history. Such documentation could be done by comparing incident cases of disease in the target population with a register of those from the same population who have been screened. This will permit an estimate of the risk in those who have been screened and in those who failed to attend screening, and the combined effect can then be compared with the prescreening period. Where such registers have not been established, a screening history should be obtained from subjects with the disease, though this may not be reliable as many are unable to recall whether a screening test has been taken in the past. Efficient surveillance requires a system of linked records. A population register (or available substitute) allows periodic call-back for rescreening at appropriate intervals. The screening program register, when linked with a disease register, permits the active surveillance of those detected with abnormalities to ensure recall for diagnosis and therapy. Evaluation of the program can then be performed with regard to assessment of:

- Management of those screened with positive tests
- Disease diagnosed between the screening interval
- Groups missed in the target population

22.12 Responsibility for Screening

Responsibility for the efficient management of screening in a country (or region) should preferably be placed on a designated official within a relevant organization. If disease control has been designated to a special agency in a country, this official should have an appointment within that agency. Alternatively, it would be appropriate to designate an official within the Ministry of Health.

In general, it is not appropriate for the director of the department or laboratory performing the tests to be given responsibility for the overall direction of the whole program. This is because the responsibilities for direction of the program are far wider than reading the screening test, but cover aspects from the identification of the target group, through their recruitment, screening, and management of abnormalities found to evaluation of the impact of the program.

The director should recognize that successful screening programs:

- Are organized as public health disease control programs, and not simply as services for providing clinical investigation
- Target the age groups at greatest and most immediate risk, concentrating on those who have never had a test
- Use population registers
- Have someone in charge who is named, has a telephone number, and can be held to account

The skills necessary to encompass the responsibility for running a successful program include epidemiology, public health, and management.

22.13 Programs

The principles that underlie the evaluation of screening programs are that they should be compatible with the program objectives; there should be standardization of nomenclature, procedures, and measurements to facilitate evaluation; and that the programs should facilitate relevant research.

The *general objectives* of screening programs are:

- To reduce the incidence of the disease (if a disease precursor is detected)
- To reduce mortality from the disease

The *specific objectives* to achieve the general objectives are:

- To provision accessible and acceptable screening services
- To recruit eligible subjects, ensuring that those at high risk are screened
- To ensure adherence to recommended screening schedules
- To ensure quality in taking the test and reading it
- To effectively communicate of results
- To appropriately follow-up if there is an abnormal test
- To provide efficient and effective treatment services

The *components* of program evaluation should be designed to be compatible with these objectives. They should include:

- Population based information systems
- Quality control systems, both internal and external

In the absence of population-based information systems, specific surveys may provide some of the needed information. They could include surveys of awareness of the need for screening in eligible and especially high-risk subjects and of the practice of screening by the target group and the advice rendered by their physicians.

Specially designed epidemiological studies, most economically of the case-control type, can also be used for program evaluation, but should in general be part of a specific research project and subject to peer review.

The data requirements for efficient program evaluation include data on the target population, and a register of all screening tests performed, with identification as to first or repeat (these records must be capable of being linked to provide a longitudinal screening history of each screenee. This requires capturing information on all changes of name and preferably using unique personal identifying numbers).

Other requirements include:

- A separate register of all abnormal tests, with data on follow-up, management, and outcome
- Data on all preclinical lesions diagnosed, classified by the recommended terminology
- Data on all cases of the disease diagnosed
- Data on all deaths from the disease
- Information on cost/manpower items, relevant to every aspect of screening

For evaluation and monitoring purposes, the data in information systems must be maintained in individually identifiable and linkable form, and the system should be so designed that it is accessible for evaluation and monitoring purposes and research as well as for the requirements of routine functioning of the program. The information on the target population should be provided by age, geographical area, and if possible the family physician of those eligible for screening. This will permit the identification of those in the population who are not being screened, who are likely to include those at highest risk, for whom special efforts will be needed to bring them into the program. Such linkages will enable the success of the measures taken to recruit eligible subjects to be determined.

The *criteria* for success or failure of screening programs relate to the objectives of the program. These include:

- Increase in awareness of the need for screening in the at-risk population
- Increase in participation rates for screening
- Reach members of the target group who have never been screened
- Improve the performance of test centers or laboratories
- Reduce the utilization of unnecessary medical procedures
- Reduce the incidence of and mortality from the disease

Thus, evaluation should include not only outcome measures, most importantly incidence and mortality from the disease, but also process measures so that if a program is not effective, the corrective actions needed can be taken.

There has to be some designated authority who will accept responsibility for program evaluation, assess the success in achieving the program objectives, and ensure corrective action is taken as necessary. Although this may vary by program and country, the primary responsibility should usually be assumed by government, though government may seek the assistance of non-government organizations in reaching the objectives. It would be desirable for government to appoint a broadly based advisory committee, with expert members drawn from a variety of disciplines. Acting on the advice of this committee, government may be able to delegate to various organizations responsible for different program components the responsibility for any corrective action necessary.

22.14 Information Systems for Screening Programs

A population-based information system is the basic building block of organized screening programs. Such information systems must be capable of supporting a diversity of goals and objectives (see below) including individual information retrieval and sophisticated aggregate and comparative data analyzes.

The range of goals and objectives to which information systems can contribute attest to their importance in assisting screening programs to achieve positive health outcomes. Information systems for screening programs should be designed to:

(a) Identify the target population
(b) Identify individuals in the target population
(c) Permit the individuals in the target population to be sent letters to:
    1. Remind her/him to attend for screening once she/he reaches the recommended age
    2. Remind her/him to reattend for screening at the recommended intervals
    3. Remind her/him to attend the local diagnosis center if an abnormality is discovered on self-examination or physician examination
(d) Monitor that action has been taken following the discovery of an abnormality
(e) Provide long-term follow-up for patients who have received treatment following the diagnosis of disease
(f) Permit linkage of individual screens at the individual level
(g) Permit evaluation and monitoring of the total system

Development of information systems will be facilitated by the introduction of permanent individual health-care identifiers. However, establishment of data bases to support screening programs need not be dependent upon unique individual identifiers and should not be delayed where such identifiers are not yet in use.

22.14.1 Goals and Objectives of Information Systems

The *goals* of an information system are:

- To facilitate enrolling the at-risk population
  For a screening program, it is essential that the data base incorporate data on the entire target population.
- To maintain information
  Information on the screening history of each screenee must be maintained on the data base; in addition, information must be organized and the data elements defined to facilitate analysis and planning.
- To provide the means for follow-up
  The information system must support communication with individuals concerning test results, the need for screening, rescreening or medical follow-up.
- To support evaluation of quality assurance

Design of the information system must incorporate the capacity for qualitative assessment of the program as a whole.

- To track utilization of screening by the target population
  A critical measure of the value of the program's information system will be its utility to follow patterns of use to determine levels of utilization.
- To monitor compliance
  Compliance with recommended screening and appropriate follow-up must be monitored by the information system to assist in evaluating the success of the overall program.

The information system developed to support the screening program must be designed to meet the following *objectives (action list)*:

- To locate the unscreened and underscreened
- To provide data to aid programs to reach special targeted groups such as the elderly
- To record detected abnormalities
- To assist in follow-up and treatment
- To assist in follow-up communications to the target population
- To support a schedule of testing
- To evaluate compliance
- To define high-risk groups
- To facilitate evaluation and planning
- To determine the cost-effectiveness of the screening program

22.14.2 Information System Design

Information system design should not be regarded as simply a technical exercise involving systems experts but must incorporate the views and data requirements of all groups involved in the program. In other words, wide-ranging consultation and participatory planning is essential. Design of information systems requires collaboration between many disciplines, the clinicians concerned with administering and interpreting screening tests, and the specialists in computer systems that can ensure that the data required to monitor and evaluate a screening program are collected, coded, collated, and made available for analysis. This process is often very difficult in developing countries. A proposal for a cervical screening information system for developing countries has been published (Marrett et al. 2002).

Without appropriate interaction on system design, opportunities may be overlooked to structure systems in ways which can serve ongoing program needs. For example, data concerning screening facilities (location, hours, wheelchair accessibility) might be identified as necessary by consumer advocates.

Participatory systems planning and development should include all stakeholders in the success of a screening program. The process of developing an information system should be sensitive to the needs of diverse interest groups. Among the groups which should be included are:

- Consumers – including representatives of health groups and all minority groups (ethnic minorities, the disabled);
- Professional colleges and associations
- The relevant non-governmental organization and the non-communicable disease (NCD) control committee
- Government including representation from public health units, health agencies, and health promotion agencies
- The research communities – including epidemiologists and academic researchers in health-care and public policy.

22.14.3 Policy Issues and Data Use

Policymakers should understand the need for the information system. In brief, the system is justified because it enables the program to be run on a daily basis, permits its performance to be assessed, and enables any necessary improvements to be made while the system can be used for evaluation and other scientific purposes, i.e., in research. The uses to which data are put are of just as much concern as is the design of the systems from which information is derived. A central concern of custodians of health-care data bases is the question of access to information and the protection of individual privacy. Extensive consultation is necessary to ensure that personal privacy is protected and that appropriate protocols are developed for the use of linked and aggregated data. In addition, other policy issues will arise that will require similar consultation and review.

22.14.4 Components of Information Systems

Information systems developed in support of health programs usually contain a number of common elements and frequently include the target population file, a registration file, data linkage capabilities (e.g., for linkage to disease registries, vital statistics systems), and mechanisms for monitoring and evaluation. Despite the likelihood of a substantial commonality of data elements when similar programs are developed in different jurisdictions, opportunities to improve program evaluation and delivery may be lost if efforts are not made to coordinate data definitions and standards.

To obtain maximum benefit from the introduction of a screening program, it is important that common definition of data elements be achieved, through appropriate consultation.

22.14.5 Conclusion on Information Systems

As part of the process of consultation needed to set up information systems for screening, a range of information management issues will have to be addressed. It is anticipated that the consultative policy and planning process advocated will provide the mechanism for addressing them.

Implementation of well-designed and monitored information systems can enhance the benefits of an organized nationwide screening program. They can help to ensure quality control by linking testing and treatment with outcomes, increase efficiency, identify underscreening of some risk groups (e.g., the elderly), support program evaluation, and help in answering research questions. These and other results can be fed back to yield further program improvement.

22.15 Conclusions

There are a number of fundamental issues that have to be resolved when considering disease control by screening. The general principles that govern the introduction of screening programs include:

- The disease should be an important health problem.
- The disease should have a detectable preclinical phase.
- The natural history of the lesions identified by screening should be known.
- There should be an effective treatment for such lesions.
- The screening test should be acceptable and safe.

The other issues range from ethics to economics. Critical issues include the population to be included in screening programs and whether or not it is possible to introduce an organized screening program. It cannot necessarily be assumed that a screening program will benefit the population to which it is applied. Not only do ethics demand that only programs with proven effectiveness be widely disseminated, it is also necessary to ensure that the program is continually monitored to confirm that effectiveness is maintained. Further, the benefits derived from the program must be clearly shown to exceed the costs, both in terms of ill health induced by the test and accompanying procedures and in economic terms. This requires that all programs are evaluated to ensure that they do meet their objectives.

In spite of these caveats, screening carries the potential for a fairly rapid and important impact on mortality from the disease, often exceeding what can currently be anticipated from other approaches to disease control, hence the continuing interest in and expectation from existing and potential programs.

References

Anderson GH (1985) Cervical cytology. In: Miller AB (ed) Screening for cancer. Academic, Orlando, pp 87–103

Andriole GL, Grubb RL, Buys SS, Chia D, Church TR, Fouad MN, Gelmann EP, Kvale PA, Reding DJ, Weissfeld JL, Yokochi LA, Crawford ED, O'Brien B, Clapp JD, Rathmell JM, Riley TL, Hayes RB, Kramer BS, Izmirlian G, Miller AB, Pinsky PF, Prorok PC, Gohagan JK, Berg CD for the PLCO Project Team (2009) Mortality results from a randomized prostate-cancer screening trial. New Eng J Med 360:1310–1319

Bach PB, Jett JR, Pastorino U, Tockman MS, Swensen SJ, Begg CB (2007) Computed tomography screening and lung cancer outcomes. JAMA 297:953–961

Baines CJ, McFarlane DV, Miller AB (1988) Sensitivity and specificity for first screen mammography in 15 NBSS centres. J Can Assoc Rad 39:273–276

Baines CJ, Miller AB, Bassett AA (1989) Physical examination. Evaluation of its role as a single screening modality in the Canadian National Breast Screening Study. Cancer 63:160–166

Bassett AA (1985) Physical examination of the breast and breast self-examination. In: Miller AB (ed) Screening for cancer. Academic, Orlando, pp 271–291

Beahrs OH, Shapiro S, Smart C (1979) Report of the working group to review the National Cancer Institute, American Cancer Society Breast Cancer Detection Demonstration Projects. J Natl Cancer Inst 62:640–709

Boyd NF, Wolfson C, Moskowitz M, Carlile T, Petitclerc M, Ferri HA, Fishell E, Gregoire A, Kiernan M, Longley JD, Simor IS, Miller AB (1982) Observer variation in the interpretation of Xeromammograms. J Natl Cancer Inst 68:357–363

Boyes DA, Morrison B, Knox EG, Draper GJ, Miller AB (1982) A cohort study of cervical cancer in British Columbia. Clin Invest Med 5:1–29

Brookmeyer R, Day NE, Moss S (1986) Case-control studies for estimation of the natural history of preclinical disease from screening data. Stat Med 5:127–138

Chamberlain J, Miller AB (eds) (1988) Screening for gastrointestinal cancer. Hans Huber, Toronto

Chodak GW, Schoenberg HW (1989) Progress and problems in screening for carcinoma of the prostate. World J Surg 13:60–64

Commission on Chronic Illness (1957) Chronic Illness in the United States: prevention of Chronic Illness. Harvard University Press, Cambridge

Cuckle HS, Wald NJ (1984) Principles of screening. In: Wald NY (ed) Antenatal and neonatal screening. Oxford University Press, Oxford

Day NE (1985) Estimating the sensitivity of a screening test. J Epidemiol Community Health 39:364–366

Day NE, Williams DRR, Khaw KT (1989) Breast cancer screening programmes: the development of a monitoring and evaluation system. Br J Cancer 59:954–958

Dillner J, Rebolj M, Birembaut P, Petry K-U, Szarewski U, Munk C, de Sanjose S, Naucler P, Lloveras B, Kjaer S, Cuzick J, van Ballegooijen M, Clavel C, Iftner T (2008) Long term predictive values of cytology and human papillomavirus testing in cervical cancer screening: joint European cohort study. Br Med J 377:a1754. doi:10.1136/bmj.a1754

Draisma G, Boer R, Otto SJ, van der Cruijsen IW, Damhuis RAM, Schroder FH, de Koning HJ (2003) Lead times and overdetection due to prostate-specific antigen screening: estimates from the european randomized study of screening for prostate cancer. J Natl Cancer Inst 95: 868–878

Eddy DM, Hasselblad V, McGivney W, Hendee W (1988) The value of mammography screening in women under age 50 years. JAMA 259:1512–1519

European Society for Mastology (1993) Report of the European Society for Mastology Breast Cancer Screening Evaluation Committee. Consensus conference on breast cancer screening. European Society for Mastology

Gastrin G, Miller AB, To T, Aronson KJ, Wall C, Hakama M, Louhivuori K, Pukkala E (1994) Incidence and mortality from breast cancer in the Mama program for breast screening in Finland, 1973–1986. Cancer 73:2168–2174

Goin JE, Haberman JD (1982) Comments on the logistic function in ROC analysis: applications to breast cancer detection. Method Inf Med 21:26–30

Hakama M (1982) Trends in the incidence of cervical cancer in the Nordic countries. In: Magnus K (ed) Trends in cancer incidence. Hemisphere Publishing, Washington, pp 279–292

Hakama M (1984) Selective screening by risk groups. In: Prorok PC, Miller AB (eds) Screening for cancer. UICC Technical report series, vol 78. International Union Against Cancer, Geneva, pp 71–79

Hakama M, Chamberlain J, Day NE, Miller AB, Prorok PC (1985) Evaluation of screening programmes for gynaecological cancer. Br J Cancer 52:669–673

Hakama M, Auvinen A, Day NE, Miller AB (2007) Sensitivity in cancer screening. J Med Screen 14:74–77

Harvey BJ, Miller AB, Baines CJ, Corey PN (1997) Effect of breast self-examination techniques on the risk of death from breast cancer. Can Med Assoc J 157:1205–1212

Henschke CI, McCaulay DI, Yankelevitz DF, Naidich DP, McGuinness G, Miettinen OS, Libby DM, Pasmantier MW, Koizumi J, Altorki NK, Smith JP (1999) Early lung cancer action project: overall design and findings from baseline screening. Lancet 354:99–105

Holowaty P, Miller AB, Rohan T, To T (1999) The natural history of dysplasia of the uterine cervix. J Natl Cancer Inst 91:252–258

Howard J (1987) Using mammography for cancer control: an unrealized potential. Cancer 37: 33–48

Howlett RI, Miller AB, Pasut G, Mai V (2009) Defining a strategy to evaluate cervical cancer prevention and early detection in the era of HPV vaccination. Prev Med doi:10.1016/j.ypmed.2008.12.022

IARC (2002) Breast cancer screening. IARC handbooks on cancer prevention, vol 8. International Agency for Research on Cancer, Lyon

Jonsson E, Hakansson S, Tabar L (1988) Cost of mammography screening for breast cancer: experiences from Sweden. In: Day NE, Miller AB (eds) Screening for breast cancer. Hans Huber, Toronto, pp 113–115

Krahn MD, Mahoney JE, Eckman MH, Trachtenberg J, Pauker SG, Detsky AS (1994) Screening for prostate cancer: a decision analytic view. JAMA 272:781–786

Mandel JS, Church TR, Ederer F, Bond JH (1999) Colorectal cancer mortality: effectiveness of biennial screening for fecal occult blood. J Natl Cancer Inst 91:434–437

Mandel JS, Church TR, Bond JH, Ederer F, Geisser MS, Mongin SJ, Snover DC, Schuman LM (2000) The effect of fecal occult-blood screening on the incidence of colorectal cancer. N Engl J Med 343:1603–1607

Marrett LD, Robles S, Ashbury FD, Green B, Goel V, Luciani S (2002) A proposal for cervical screening information systems in developing countries. Int J Cancer 102:293–299

McLelland R, Pisano ED (1992) The politics of mammography. Radiol Clin North Am 30:235–241

Miller AB (ed) (1978) Screening in cancer. A report of the UICC international workshop in Toronto. UICC Technical report series, vol 40. International Union Against Cancer, Geneva

Miller AB (1981) An evaluation of population screening for cervical cancer. In: Koss LG, Coleman DV (eds) Advances in clinical cytology. Butterworths, London, pp 64–94

Miller AB (1987) Early detection of breast cancer. In: Harris JR, Henderson IC, Hellman S, Kinne DW (eds) Breast diseases. Lippincott, Philadelphia, pp 122–134

Miller AB (1991) Issues in screening for prostate cancer. In: Miller AB, Chamberlain J, Day NE, Hakama M, Prorok PC (eds) Cancer screening. Cambridge University Press, Cambridge, pp 289–293

Miller AB (1994) Screening for cancer: Is it time for a paradigm shift? Ann R Coll Physician Surg Can 27:353–355

Miller AB (1996) Fundamental issues in screening for cancer. In: Schottenfeld D, Fraumeni JF Jr. (eds) Cancer epidemiology and prevention, 2nd edn. Oxford University Press, New York/Oxford, pp 1433–1452

Miller AB (2004) Natural history of cervical cancer. In: Rohan TE, Shah K (eds) Cervical cancer, from etiology to prevention. Kluwer Academic Publishers, Dordrecht, pp 61–78

Miller AB, Tsechkovski M (1987) Imaging technologies in breast cancer control: summary report of a world health organization meeting. AJR 148:1093–1094

Miller AB, Lindsay J, Hill GB (1976) Mortality from cancer of the uterus in Canada and its relationship to screening for cancer of the cervix. Int J Cancer 17:602–612

Miller AB, Chamberlain J, Day NE, Hakama M, Prorok PC (1990) Report on a workshop of the UICC project on evaluation of screening for cancer. Int J Cancer 46:761–769

Miller AB, Baines CJ, Turnbull C (1991a) The role of the nurse-examiner in the National Breast Screening Study. Can J Public Health 82:162–167

Miller AB, Anderson G, Brisson J, Laidlaw J, Le Pitre N, Malcolmson P, Mirwaldt P, Stuart G, Sullivan W (1991b) Report of a national workshop on screening for cancer of the cervix. Can Med Assoc J 145:1301–1325

Miller AB, Baines CJ, To T, Wall C (1992) Canadian national breast screening study: 2. Breast cancer detection and death rates among women age 50–59 years. Can Med Assoc J 147: 1477–1488

Miller AB, Nazeer S, Fonn S, Brandup-Lukanow A, Rehman R, Cronje H, Sankaranarayanan R, Koroltchouk V, Syrjanen K, Singer A, Onsrud M (2000a) Report on consensus conference on cervical cancer screening and management. Int J Cancer 86:440–447

Miller AB, To T, Baines CJ, Wall C (2000b) Canadian National Breast Screening Study-2: 13-year results of a randomized trial in women age 50–59 years. J Natl Cancer Inst 92:1490–1499

Miller AB, Madalinska JB, Church T, Crawford D, Essink-Bot ML, Goel V, de Koning HJ, Maatanem L, Pentikainen T (2001) Review: health-related quality of life and cost-effectiveness studies in the European randomised study of screening for prostate cancer and the U.S. Prostate, Lung, Colon and Ovary trial. Eur J Cancer 37:2154–2160

Miller AB, To T, Baines CJ, Wall C (2002) The Canadian National Breast Screening Study-1: breast cancer mortality after 11 to 16 years of follow-up. A randomized screening trial of mammography in women age 40 to 49 years. Ann Intern Med 137:305–312

Morrison AS (1985) Screening in chronic disease. Oxford University Press, Oxford, pp 48–63

Morrison AS, Brisson J, Khalid N (1988) Breast cancer incidence and mortality in the breast cancer detection demonstration project. J Natl Cancer Inst 80:1540–1547

Moskowitz M, Pemmaraju S, Fidler JA, Sutorius DJ, Russell P, Scheinok P, Holle J (1976) On the diagnosis of minimal breast cancer in a screenee population. Cancer 37:2543–2552

Moss SM (1991) Case-control studies of screening. Int J Epidemiol 20:1–6

Moss S, Draper GJ, Hardcastle JD, Chamberlain J (1987) Calculation of sample size in trials of screening for early diagnosis of disease. Int J Epidemiol 16:104–110

Moss SM, Cuckle H, Evans A, Johns L, Waller M, Bobrow L, Trial Management Group (2006) Effect of mammographic screening from age 40 years on breast cancer mortality at 10 years' follow-up: a randomised controlled trial. Lancet 368:2053–2060

Nieminen P, Kallio M, Anttila A, Hakama M (1999) Organised vs. spontaneous Pap-smear screening for cervical cancer: a case-control study. Int J Cancer 83:55–58

Prorok PC, Chamberlain J, Day NE, Hakama M, Miller AB (1984) UICC workshop on the evaluation of screening programmes for cancer. Int J Cancer 34:1–4

Sasco AJ, Day NE, Walter SD (1986) Case-control studies for the evaluation of screening. J Chronic Dis 39:399–405

Schechter MT, Miller AB, Baines CJ, Howe GR (1986) Selection of women at high risk of breast cancer for initial screening. J Chronic Dis 39:253–260

Semiglazov VF, Sagaidak VN, Moiseyenko VM, Mikhailov EA (1993) Study of the role of breast self-examination in the reduction of mortality from breast cancer. Eur J Cancer 29A:2039–2046

Strong K, Wald N, Miller A, Alwan A, on behalf of the WHO Consultation Group (2005) Current concepts in screening for noncommunicable disease: World Health Organization Consultation Group report on methodology of noncommunicable disease screening. J Med Screen 12:12–19

Swets JA (1979) ROC analysis applied to the evaluation of medical imaging technologies. Invest Radiol 14:109–121

Tabar L, Fagerberg CJG, Gad A, Baldetorp L, Holmberg LH, Gröntoft O, Ljungquist U, Lundstrm B, Månson JC, Eklund G, Day NE (1985) Reduction in mortality from breast cancer after mass screening with mammography: randomized trial from the breast cancer screening working group of the Swedish National Board of Health and Welfare. Lancet i:829–832

Tabar L, Fagerberg G, Duffy SW, Day NE (1989) The Swedish two county trial of mammographic screening for breast cancer: recent results and calculation of benefit. J Epidemiol Community Health 43:107–114

Task Force (1976) Cervical cancer screening programs. The Walton report. Can Med Assoc J 114:1003–1033

Task Force (1982) Cervical cancer screening programs: summary of the 1982 Canadian task force report. Can Med Assoc J 127:581–589

UK Trial of Early Detection of Breast Cancer Group (1988) First results on mortality reduction in the UK Trial of Early Detection of Breast Cancer. Lancet ii:411–416

Walter SD, Day NE (1983) Estimation of the duration of a pre-clinical disease state using screening data. Am J Epidemiol 118:865–886

Walter SD, Stitt LW (1987) Evaluating the survival of cancer cases detected by screening. Stat Med 6:885–900

Weiss NS (1983) Control definition in case-control studies of the efficacy of screening and diagnostic testing. Am J Epidemiol 116:457–460

Weiss CH (1998) Evaluation. Prentice Hall, Englewood Cliffs

Wilson JMG, Junger G (1968) Principles and practice of screening for disease. World Health Organization, Geneva

Workshop Group (1989) Reducing deaths from breast cancer in Canada. Can Med Assoc J 141:199–201

Epidemiology in Developing Countries 23

Klaus Krickeberg, Anita Kar, and Asit K. Chakraborty

Contents

K. Krickeberg (✉)
University of Paris V (retired), Bielefeld, Germany

A. Kar
School of Health Sciences, University of Pune, Pune, MH, India

A.K. Chakraborty
BIKALPA, Koramangala, Bangalore, India

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_39,

23.1 Introduction

Modern epidemiology can boast of a precise definition, clearly formulated basic concepts, and well-elaborated methods, all of them described elsewhere in this handbook. They are in principle the same in developing and in developed countries. What is different is the *framework*, the choice of *topics* and *applications* to be treated given local necessities, but also the battery of *tools* adapted to typical tasks and available under local conditions, and finally, the type of *difficulties* that epidemiological work is facing.

This chapter attempts to outline these differences. Still, both developing and developed countries vary enormously also between themselves. Moreover, many countries belong to one of the two categories when applying certain criteria like income but to the other one regarding other aspects like education. We will therefore not adhere throughout to a fixed definition of a developing or a developed country.

In Sect. 23.2, after a short description of essential features that are specific to health systems in developing countries, we review the main needs of these countries in the realm of epidemiology. The last four sections are devoted to the methods of acquiring and applying epidemiological knowledge. Of these, global health information systems are a particularly striking specific component of the health structures in developing countries because they practically exist only there. Sample surveys also play a more important role than in developed countries, whereas the opposite is true for more advanced epidemiological studies.

Public health and health systems in India can be used as an illustrative example of the challenges and achievements of epidemiological investigations in designing public health interventions in India. India is the second most populous country in the world with a population of over 1.2 billion in 2010 and an annual birth cohort of approximately 25 million in 2001 (Census of India 2001). This gigantic population is distributed through diverse geographical terrains and shows a wide spectrum of socioeconomic development, cultural heterogeneity, and linguistic diversity. The phenomenon of health planning is relatively recent, initiated after independence, a little over 60 years ago. The historic evolution of the health sector has witnessed the development of both public and private health sectors. There is currently no mechanism to capture health data from the private medical sector. Adding to this complexity is the fact that India has plural systems of medicine, namely, Indian systems of medicine such as ayurveda, unani, siddha, naturopathy, as well as allopathy and homeopathy. These different streams of medical practice, with different methods of diagnosis and categorization of disease, enhance the challenge of collecting accurate, reliable, and up-to-date morbidity and mortality data.

In Sect. 23.3, we first describe the organization of the health system, pointing out the huge outreach and the well-defined Health Management Information System (HMIS) that exists for regular collection of health data. We next point out the caveat to this data collection system, caused by the presence of a colossal private health sector from where there is no mechanism to capture morbidity data. We then move to the specific example of tuberculosis, firstly showing how epidemiological research, done outside the health system, has helped in shaping the tuberculosis control program. We show the impact of international "vertical" funding in strengthening the health information system and the continued role of repeated research in contributing to the understanding of the efficiency of the control program in case detection. We complete the circle of the discussion by reiterating that despite the best efforts of establishing a HMIS, accurate data for basing health interventions will be lacking in India unless the country tackles the formidable task of designing an integrated data collecting system that includes health information from both public and private health sectors.

23.2 General (by K. Krickeberg)

23.2.1 The Framework: Health in Developing Countries

The specific form of epidemiological research and applications in a given country is conditioned to a large extent by the way health care is organized and functioning. Here, we encounter many differences between developed and developing countries. The former are generally richer and can spend more money on health in absolute terms but usually also in relative terms as measured in relation to national income.

More specifically, the way curative care is organized differs substantially between the two groups. This is especially true for primary health care in the classical sense of health care that begins at the time of the first encounter between a patient and a provider of health service. In developed countries it rests essentially with the general practitioner who is a trained physician and to a lesser extent with doctors in hospitals and policlinics. In developing countries it is mostly offered by *communal health centers* (CHCs) or comparable health facilities that are mainly staffed with health workers like nurses and midwives who are less trained but sometimes more efficient than physicians. Regarding secondary health care, hospitals in the least developed countries tend to be concentrated in large cities, especially in the capital, to the detriment of the rest of the country.

General hygiene and specific preventive measures in developing countries suffer particularly from the lack of material means and knowledge. Bureaucratic weak administrative structures and lack of management training and experience also have much bearing on public health and so do insufficient civil registries and other records, especially demographic ones. It is in this environment that, for instance, the burden of AIDS has been incommensurably heavier than in developed regions of the world.

The main reason next to poverty for which the health system in developing countries has taken such a different path from that of developed ones is founded in their historical evolution and in particular in the role of contributions from outside. These were twofold.

Firstly, health care in colonies was initially built up to a large extent by missionaries and then by the colonial masters, the latter adapting it mainly to their own needs, sometimes to the detriment of indigenous medicine. This building up continued later in the form of bilateral cooperation with developed countries, especially training, and also embraced developing countries that had never been colonized like Thailand or Turkey, often in synergy with national efforts. In general it emphasized individual health care and often suffered from the concentration of hospitals in urban centers. Again, there were exceptions, some developing countries like Vietnam succeeding in making substantial progress in public health, too, and in creating networks of CHCs and small hospitals that covered the entire country.

Secondly, international programs have been a very important source of contributions from outside. They were mostly designed by WHO but implemented by organizations like UNICEF, international or national non-governmental organizations, and national health authorities or institutes like ministries of health or national hygiene institutes. They are installed and run *vertically*, from their respective top administrations down to the basic providers of health care. They tend to strengthen public health and are sometimes curative, sometimes preventive, and sometimes both. They have shaped large sectors of the health system of developing countries. Since they usually involve a lot of epidemiological work, we will list some of them here. For details, see the relevant publications of WHO.

The first major international programs after World War II concerned tuberculosis and malaria. Other programs directed against particular diseases followed, for example, against smallpox, which lead to its eradication, poliomyelitis, goiter, cataract, and finally HIV. There were some programs of more regional importance like leprosy, schistosomiasis, onchocerciasis (river blindness), meningitis, and arthropod-borne virus diseases, especially yellow fever and dengue fever.

In addition, WHO programs of a different type appeared which were not meant to prevent or cure a specific disease but to prevent death caused by certain unspecific manifestations of any of a whole group of acute ailments: CDD and ARI. The core of CDD (Control of Diarrheal Diseases) consisted in preventing death from dehydration of children under the age of 5 who suffered from acute diarrhea. The main method was *oral rehydration* regardless of the cause of the diarrhea. Hygienic measures to prevent diarrheas were included but not the main objective. The philosophy of ARI (Acute Respiratory Infections) was similar, namely, to prevent, by a non-specific antibiotic standard treatment, the death of children under 5 who had caught an acute respiratory infection like pneumonia. Preventive action against the infections themselves was a side issue.

Essential drugs is a WHO program of still a different nature, and so is reproductive health, which often boils down to plain family planning.

The most beneficial WHO program is probably EPI (Extended Program on Immunization) by which a high number of cases of severe diseases and deaths have been prevented. Its objective is the systematic vaccination of all children under 1 or 2 years of age against measles, poliomyelitis, diphtheria, tetanus, and pertussis (whooping cough), plus a dose of BCG that is, however, controversial and no longer generally recommended. Moreover, the program includes a vaccination of all women in childbearing age against neonatal tetanus. The vaccination of children against Hepatitis B was added later.

Many other international programs of varying importance exist or have existed, above all MCH (Mother and Child Care), reproductive health, but also nutrition, vitamin A deficiency, rehabilitation of war victims, etc.

In many developing countries, a plethora of international organizations is running a health program, mostly of the vertical type. These programs have unfortunately been very little coordinated with each other. Often they draw many of the resources of the health system to themselves, to the detriment of other activities, detracting in particular from the normal routine. In particular, EPI was sometimes overly stressed. The conceptual and operational bases of an integration of the various aspects of primary health care have been expounded long ago (Krickeberg 1989). A program IMCI (Integrated Management for Childhood Diseases) was launched by WHO around 1990.

We are now going to outline the role that epidemiological work plays and, more importantly, ought to play within this framework.

23.2.2 Epidemiological Needs of Developing Countries

In the area of *descriptive* epidemiology in the classical sense, that is, *health statistics*, the needs of developing countries are in principle hardly different from those of developed ones: incidence or prevalence of major diseases by sex, age, time, and place of residence plus disease-specific mortality, infant mortality, maternal mortality, etc. (cf. chapter ▸Descriptive Studies of this handbook). These statistics are one of the pillars of the *management* of the health system, especially yearly budgeting. They ought to be used also for rational planning and implementing *global health strategies*, but this aspect still leaves much to be desired apart from international vertical programs. They are often presented in official health statistics publications, for example, health yearbooks that sometimes appear to be their main justification.

The actual differences between the needs for descriptive epidemiology within the two groups of countries result from the different weight that the various categories of disease have or that health authorities attribute to them. Infectious diseases certainly weigh much more in developing countries than in developed ones in spite of their resurgence in the latter (cf. chapter ▸Infectious Disease Epidemiology of this handbook). Incidence or prevalence figures should be available regularly. Classical *epidemic surveillance* is of course of utmost importance.

Health authorities in developing countries still tend to neglect non-infectious ailments like cancer, cardiovascular diseases, diabetes, chronic respiratory diseases, and osteoarthritis. Their incidence, prevalence, and mortality are now in some of these countries of the same order of magnitude as that of the main infectious diseases. This “epidemiological transition” is partly due to the so-called demographic transition, that is, longer life, because the incidence of most of these non-infectious diseases is increasing with increasing age. It is also due to changing life-style, for example, less healthy food and increasing air pollution. Obesity is becoming prominent both as a risk factor and as a disease of its own. Hence reliable figures on the frequency of these diseases, too, are required.

Very few indicators are known on *nicotine addiction*, one of the most menacing epidemics in developing countries.

There are two more areas where the emphasis that is placed on knowing the basic indicators is different in developing countries on the one hand and developed on the other, namely, Mother and Child Care (MCH) and reproductive health.

The most *specific* requirements regarding incidence or prevalence usually emanate from the managers of special programs, mainly international vertical ones, who need them for planning, running, monitoring, and evaluating their programs.

In the domain of *analytical (observational) epidemiology*, that is, the study of risk factors, the needs of developing countries have been less well formulated than in that of health statistics. There has been a tendency to rely on results obtained in developed countries or by teams from there who found interesting study objects in a developing country but often contributed little to furthering the epidemiological capacities of their hosts. However, there is an obvious need of knowing the essential risk factors of diseases such as tuberculosis, AIDS, malaria, dengue fever, nicotine addiction, cancer, cardiovascular diseases, and diabetes in order to plan and implement preventive measures. Even when studies of such risk factors are made, rules on how to implement practical conclusions from them, especially about prevention, are not always derived in a satisfactory way.

Regarding observational epidemiology of non-infectious diseases, specific to developing countries, there is obviously an urgent need regarding, above all, the investigation of new nutritional risk factors and of environmental exposures in large cities, but also of infections, for example, viral infections as causes of cancer. Risk factors for cardiovascular diseases under the conditions of a developing country ought to be known better.

For infectious diseases, the principal risk factor being the infective agent, there has been less motivation to investigate contributing factors like genetic, social, and environmental ones. Studies have mainly centered on the pathways of the pathogen and their effect on infections including mathematical modeling. Their practical conclusions regarding, for example, hygiene are more or less known. For concrete illustrations, it suffices to think of diarrhea and AIDS. However, studies of general risk factors of the type mentioned above are indeed direly needed in developing countries. For example, in the realm of the program ARI, what is the effect of smoke in dwellings on acute respiratory diseases? Which factors influence the malaria cycle, and how?

Preventive measures are determined by the knowledge of risk factors. The principal classical preventive measures in developing countries have been on the one hand person-based *interventions* of the type treated in chapter ▸Intervention Trials of this handbook, especially *immunizations* in the framework of EPI, and, on the other hand, community-based *health promotion* as described both in chapters ▸Intervention Trials and ▸Community-Based Health Promotion of this handbook, often connected with vertical programs. As examples we can quote classical hygiene, treating bednets against anopheles, and urging the community to cooperate in a given program. Campaigns for healthier nutrition or against smoking are still in their infancy and so are screening programs along the lines of those presented in chapter ▶Screening of this handbook for developed countries. There is a need to evaluate the effect of such activities in a more precise way.

Regarding *immunizations*, there is again a tendency to make do with the results about their efficacy obtained in developed countries. Sometimes, however, the situation may be different. A case in point is the problem of the optimal period for vaccinating against *measles*, the so-called window. In developed countries, this vaccination can wait until the maternal antibodies still present in the infant's blood can no longer interfere with its immunizing action. In developing countries, however, the infectious potential around a child is often very high from the beginning. The child's maternal antibodies do not provide sufficient protection, hence early vaccination is called for. To overcome this dilemma, WHO decided to advocate the use of a higher-titer vaccine of the "Edmonston-Zagreb" type in spite of the existence of epidemiological studies, which gave rise to the fear that this might lead to a higher overall-mortality among vaccinated female children. This was indeed observed in later studies, and the vaccine had to be withdrawn (Das Gupta et al. 1997; in particular the article Aaby 1997).

Community-based interventions in developing countries of which we have mentioned above some examples are obviously much needed. The objectives are often very different from those in developed ones. Their evaluation is in principle the same, but like the actions themselves, it is often much more poorly designed and executed. Similarly, rigorous person-based intervention trials are largely lacking.

Up to now, we have dealt with *global* epidemiological needs in the sense that they concern a whole country. It is health authorities at the top of the hierarchy who ought to be aware of them. In contrast, there is what WHO once called "epidemiology at the basis," not to be confused with "field epidemiology." This is *local* epidemiology, which is tied to a smaller geographical or administrative entity, usually a commune. For example, knowing the so-called disease spectrum, that is, the relative frequencies of the various diseases including traumata which appear at a CHC, has not only a practical value for managing the center but also an educative one. It may be a motivation for the health worker as well. The same holds for other simple activities in local observational epidemiology like linking an incidence to a social factor or a geographical one, for example, stagnant waters to shigellosis. While nurses often design ingenious charts, maps, and other devices to monitor certain epidemiological features of their community including rules how to apply them, a general and systematic approach is wanting.

In *experimental* epidemiology including intervention activities, the needs of developing countries are manifold. They arise in all branches of that area. There is *clinical* epidemiology both in the form of studies in diagnostics and of clinical trials of curative treatments, and there are trials to evaluate preventive actions. All of them play a particular role in vertical programs. Let us elaborate a bit and look at a few examples.

As said above, diagnostics and curative treatment in developing countries are often very different from what is customary in the developed world. To a large extent they take place in CHCs and are performed by only partly trained health workers under poor material conditions, in particular in the absence of laboratory equipment. Therefore, simplified and *standardized decision rules* for case management have been designed.

For example, diagnoses have usually to be made on the basis of only some clinical symptoms without any laboratory analysis. In regions of Cambodia where malaria is endemic, the symptoms "fever" and "headache" used to entail the preliminary diagnosis "malaria." Similarly, a differential diagnosis of amoebic and bacterial dysentery (shigellosis) must often be founded on the clinical symptoms "headache," "fever," "blood in the stool," and "mucus in the stool." In both cases, a more reliable diagnosis may take time because a blood or stool analysis can only be done in a laboratory that is far away, or it may not be possible at all. However, it is usually both necessary and common to treat the patient at once, having only a preliminary diagnosis at one's disposal. If one wants to judge the merits and dangers of this strategy, one needs to know the epidemiological characteristics of the underlying preliminary diagnostic decision rule, that is, its *sensitivity*, *specificity*, and *prognostic values*. One may also wish to compare this clinical decision rule with alternatives of the same type in order to select the best one.

The program CDD is another example. Here, the result of the diagnosis is not expressed in the form of a particular disease but as a degree of dehydration, either 0, I, II, or III, determined as a simple function of a few easily recognizable clinical symptoms. The treatment, too, has a standardized form and depends only on this degree, in particular oral rehydration in case II. Again, the epidemiological characteristics of the diagnostic rule need to be known in order to evaluate the usefulness of the entire strategy. ARI works analogously.

The preceding discussion of diagnosis applies in a similar way to simple standardized *curative treatments* that are being used widely in developing countries. Their efficacy ought to be estimated by appropriate clinical trials.

Traditional medicine, too, is an area where the epidemiological necessities of developing and developed countries obviously differ. In order to bridge the gap between traditional curative treatments and those taught in medical schools at universities and to integrate the useful part of the former into the health system of developing countries, clinical trials including meta-analyses are required. There is in principle no difference between the two curative systems. After all, much of the "Western" pharmacopoeia had its origin in traditional medicinal plants, for example, quinine, aspirin, digoxin, and, more recently, viagra. Unfortunately, until not long ago, medical herbs and animals that were known for centuries if not millennia have been studied only from a purely pharmacological and chemical point of view, to extract the "active principle," without raising the question of their actual efficacy. It is only recently that traditional cures of any kind have been subjected to epidemiological scrutiny, first in India and China, and then in developed countries, too.

Many of the deficiencies mentioned before result from poor training of physicians and health officials. Good *teaching* in the area of epidemiology is indeed one of the most basic needs of developing countries. Many courses on particular subjects are being organized by international organizations, often under the heading "field epidemiology," but the usual epidemiological curricula in universities both for medical students and for future public health specialists are usually very weak. They do not convey a global view on the fundamental ideas.

For example, few physicians and health officials will realize that there is a fundamental difference between the epidemiology of infectious and non-infectious diseases in addition to the obvious ones such as the role of the infectious agents and of epidemic surveillance. It consists in the *indirect* effect of not only preventive but also of curative measures against infectious diseases; the latter is absent for non-infectious ones. Treatment of cases of an infectious disease usually reduces the number of sources of infection and thereby contributes to lower the incidence. Analyzing and predicting this indirect effect requires mathematical modeling (chapter ▸Infectious Disease Epidemiology of this handbook). In developing countries this was done for various diseases and recently also attempted for AIDS. However, the lack of theoretically well-trained epidemiologist in developing countries rarely permits to organize such analyses that take into account the specific situation there.

We now turn to the question of how and to which extent the epidemiological needs of developing countries can be satisfied in theory and practice. We will first describe the most important tool, namely, health information systems and its role in epidemiology. This role is fundamental but is often not clearly recognized; it is therefore treated in some detail. In the last two sections, we examine the contributions of the two other tools that exist, namely, health surveys and epidemiological studies.

23.2.3 Health Information Systems in Developing Countries

A health information system (HIS) links different institutions that are concerned with health between each other. It consists of mechanisms to *collect*, *transmit*, *analyze*, and *exploit* information on health in the widest sense. This is mostly done regularly, in a *routine* fashion, but may be supplemented by an ad hoc exchange of information. For example, most hospitals have their own, sometimes computerized, system by which information flows between its various wards and administrative units.

Here, we will be dealing only with HISs which go beyond a single institution and where information is transmitted between CHCs, policlinics, hospitals, and public health institutions and administrations. For instance, a central disease register as described in chapter ▸Use of Health Registers of this handbook amounts to a HIS where reports from physicians, hospitals, and laboratories are directed to the office where the data are stored and handled.

We are now going to restrict our scope further by focusing on much larger HISs that may deal with many sorts of information and link institutions of very different nature. In the extreme case, such a system may be founded on *regular reports* on *all* major events and activities from *all* health institutions of a *whole country* to higher health authorities, culminating in the Ministry of Health. The bulk of information emanates from the basic health institutions that are in direct contact with the population for curative or preventive activities: private practitioners, health centers, policlinics, small hospitals, individual wards of large hospitals, hygiene teams, etc.

Developing countries have, as a rule, started very early to build HISs, sometimes rudimentary, sometimes quite elaborate. They were motivated by the need to manage scarce resources efficiently, including traditional budgeting, and to control the functioning of the various components of the health system. Epidemiological information for planning and implementing health strategies played originally a secondary role apart from simple health statistics based on statistics of treatments, but over the years HISs have become an essential and often the main source of epidemiological knowledge.

In spite of the necessity felt by health planners to be equipped with a reliable HIS, the systems actually constructed have almost consistently been suffering from serious deficiencies. In some countries, for example, in Africa and Central America, they did not reach far enough into the countryside nor did they cover the essential subject matter well enough. In others, on the contrary, especially in socialist countries, they were usually too heavy and bureaucratic, attempting to cover everything and containing many redundant or superfluous elements that impeded their functioning. Neither the former nor the latter were designed following clear ideas and guidelines including a rational logical structure.

In the beginning, HISs were installed by the central health authorities, usually the Ministry of Health. They tried to build a more or less centralized HIS under which basic health institutions had to send routine reports on their general activities following a hierarchical path, for example, from CHCs to district health administrations, from there to provincial health authorities, and ending up in the Ministry. Unfortunately, given the deficiencies of such systems mentioned above, in some countries other central institutions had to build their own HIS that ran parallel to that of the Ministry. For example, a national hygiene institute would set up a separate reporting system on *infectious diseases* including *epidemic surveillance*. Central tuberculosis or malaria institutes would establish similar systems of their own. All of this caused a lot of extra work in basic institutions, especially in CHSs.

The worst confusion, however, resulted from the management of international programs. As said above, managers usually felt the need for a centralized and *vertical* approach, working from the top to the bottom. This involved specific systems for providing the required information in the form of basic demographic, epidemiological, sociological and economic indicators, and of indicators to be updated all the time, for instance, on the logistics of drugs or on vaccination coverage. Thus a plethora of information systems emerged for the various programs. As a rule these manifold HISs were coordinated neither with each other nor with the system of the Ministry and other existing systems. The burden of writing and filing all the required reports rested mainly with the local health workers who were naturally more interested in their purely medical activities. It became frequently unbearable. The situation in Vietnam as described in a report (Krickeberg 1999) is a good example though there are great differences between countries.

The question of how to design a *single* information system that could be handled efficiently and at the same time serve all the essential needs of health care, that is, a so-called *integrated* HIS, has therefore been discussed occasionally. Although some developing countries have in fact set up reasonable HISs, for example, Cuba, the principles on which such systems should rest have been investigated only recently in a systematic way (Krickeberg 1994, 2007; Lippeveld et al. 2000). The remaining part of this section will be devoted to a short presentation of these principles, starting with the functions of HISs, passing on to the properties which they should have in order to do what we want them to do, and finally looking at their structure. More details can be found in the three references just quoted.

Regarding functions, *budgeting* and all *health management* that go beyond the smallest entities of the health system require a good HIS. This holds for the general standing health system as well as for special vertical programs that rely, for instance, on an efficient logistics for drugs or vaccines. *Health insurance* could in principle get from an integrated HIS all the information it needs for planning and running. An example of an application would be to estimate the average cost caused by a case of a particular disease. In the context of this handbook, though, we are only dealing with the epidemiological functions of a HIS, to be described in the following section.

Regarding desirable properties, they are largely dictated by experience. The system should be *integrated*, that is, cover all health institutions and all kind of numerical information. In particular, the medical and epidemiological aspects must not be separated from the economic and managerial components. So-called Health Management Information Systems usually cover *some* clinical and epidemiological information but not enough. They are inadequate in most cases, and the term is misleading.

The HIS ought to have a *transparent*, *logical structure*. This is not only indispensable for designing and running the system efficiently but also constitutes, as again many experiences have shown, a powerful motivation for the health workers who handle it. Moreover, only a well-structured HIS can be computerized although there have been quite a few, always futile, attempts at saving a bad system by using computers.

The system needs to be shaped in view of *well-defined objectives* and incorporate rules how to analyze and exploit the information obtained in order to reach these objectives.

Given its objectives, it has to be *minimal*, free of superfluous elements and in particular of redundancies, and has to function at minimal cost. This has, in particular, an implication regarding the *flow of information* between institutions: routine reports must be filed only from and to institutions as really needed and following clear operational priorities.

The system should be *flexible* so it can be adapted to changes of all kind, in particular to varying epidemiological situations, but also to alterations of the structure of the health system or of health strategies and medical knowledge and techniques, economic and social changes of the country, and new information technology.

The information obtained must be as *reliable* as one can realistically hope for.

There have to be clearly defined rules indicating what to do about *wrong* or *missing information*, and there have to exist *error-correcting* procedures along the lines of chapters ▸Measurement Error and ▸Missing Data of this handbook; they must of course be sufficiently simple to be applied to every single report. We will come back to this at the end of the present section and in the next one.

Moving on to the underlying structural and technical principles, we have to recall first a few elementary definitions because there is no general agreement about them. As always in statistics when talking about information in a concrete context, we need to specify first a *population* (set) of *units* (elements). It is sometimes called the "target population" and need of course not consist of humans. Examples might be all children under 5 living at a given moment in a given village, or all consultations done during a certain month in a certain health center.

Information has either the form of *data* or of *indicators*. Data are the values of a *variable*, which is a function defined on a population, like the function *age* which assigns, to each child, his or her age or the function *diagnosis* that attributes to every consultation the diagnosis made, properly coded. A *register* is a concrete, explicit representation of the data of one or several variables as described in chapter ▸Use of Health Registers of this handbook, usually on paper or on a computer disk. We will not enter into technical details like sub-registers or linked registers.

An *indicator*, in contrast to data, concerns a population as a whole but not individual units. It depends on one or several variables defined on the same population. Two examples, concerning our first and second example of a variable above: the mean age of the children and the number of "diarrhea" cases recorded. An example of an indicator depending on two variables, again for consultations as units: the mean age of the children under 5 who consulted for "diarrhea." In practice, an indicator is computed from *all* the values of the underlying variables as given in a register. Let us note in passing that WHO has given a completely different definition of an indicator, which is at variance with the definition employed by statisticians.

Registers are *the* fundamental, and often neglected, components of a HIS. They need to obey the following "golden" rule: in any basic institution, there is *only one register for a given population*, that is, for a given type of unit. For example, it is compulsory that there exist, in a health center or a ward of a hospital, only one register of consultations, not a "general" one for the Ministry, and one more for each vertical program and for other special purposes. One reason is simplicity, which in turn is a strong motivation for the health worker to use the system conscientiously. The main reason, however, is epidemiological and will be discussed in the following section.

Well-designed registers are in the first place the main tools for the *local use* of the HIS, both clinical and epidemiological as defined in the preceding section and to be taken up again in the following one. Their multiple functions demand a clear definition of the variables involved.

The second function of a register, namely, to serve as the basis for the routine reports to be filed, should be determined by rigorous structural rules as well. This applies both to paper-based and computer-based HISs. We note first that most information contained in routine reports is in the form of *indicators*, be they epidemiological, economic, or others. The transmission of *data* on individual units, for example, on cases, is indeed rare and mainly restricted to epidemic surveillance and registers of special diseases. For indicators resulting from clinical activities which are the most prominent ones, the structural rule in question looks like this: the clinical act giving rise to the data, the immediate use of these data in the act, the filling in of the relevant register, and the drawing up of the reports based on these data have to be *integrated conceptually* and *technically*. In particular, the *layout* of the register and that of the reporting forms need to be closely coordinated so as to allow calculating the indicators and writing the report in a single, transparent, and almost automatic operation.

In this way, many indicators of quite a different nature can be easily computed and transmitted to various institutions and according to needs that may change, all based on the *same* registers and on the *same* variables. However, on no account should there exist a register that is not tied operationally to some function within the health system and whose only purpose is to calculate and report indicators.

There have been attempts, for example, by WHO, at designing HISs by starting with a list of the indicators to be covered in view of the goals of the system. This idea may be tempting at first sight but is in fact naïve. Such lists have always been long and very much subject to debate, they leave no room for flexibility, and they are at variance with the basic structural principles as outlined above. What we have to fix in the beginning are the *variables* needed; to do this well is crucial. The second step will then be the design of the *registers*.

Reliability of the information is a particularly difficult problem in the context of a developing country. Standard procedures to assure quality such as those described in the chapters ▶Epidemiological Field Work in Population-Based Studies and ▶Quality Control and Good Epidemiological Practice of this handbook can only be applied to a rather limited extent. Motivation of the health worker is a more efficient measure, but gaps and errors will remain. The topic of what to do when confronted with them has many facets. Chapters ▶Measurement Error and ▶Missing Data of this handbook concern errors and missing data in a *single* epidemiological study. In the realm of HISs, however, we have to handle them *routinely*, for example, for every monthly report of a district health administration to its superiors. To this end, simple rules and algorithms need to be developed which can function under the specific conditions (Krickeberg 1994, 2007).

Let us now try to elucidate to which extent, and how, HISs can fulfill the epidemiological functions enumerated in Sect. 23.2.4.

23.2.4 Epidemiological Insights from Health Information Systems

It is a truism that, *potentially*, epidemiological needs can be satisfied via a HIS to the extent, and only to the extent, that the necessary data or combinations of data are recorded in the system. The practice is a bit different, but let us look at the mechanisms anyway.

The traditional basic need is classical health statistics. It amounts to obtaining regularly indicators on incidence, prevalence, and mortality for, in principle, all diseases including traumata. In developing countries, these are usually derived from a HIS. This system may contain special components in the form of "disease registries" for particular diseases such as a cancer registry; they have been set up in several countries with varying degrees of completeness. Health statistics published by Ministries of Health and WHO are essentially based on national HISs. Let us look at the mechanisms.

Existing HISs have usually been built in view of establishing health statistics and of managing health activities albeit in a rudimentary fashion, for example, yearly budgeting. We have sketched the principles for reforming older HISs or building new ones. In any case, we are facing the following questions: What kind of registers are there? Which variables are being recorded and how? How is information extracted from the registers and transmitted to its destinations?

As stated in the preceding section, a register is tied in a natural way to a particular operation that is routinely executed in the health system, above all a clinical one. Thus, the main register in a CHC is the register of *consultations*. Next to the date of the consultation and data to identify the patient, it features variables like symptoms or syndrome, sometimes a tentative diagnosis, and treatment. Recording of the relevant data serves, in the first place, the clinical act "consultation" itself. Clearly defined variables may indeed help the health worker to decide about his diagnosis and treatment as explained in Sect. 23.2.3 for CDD and ARI cases.

The main traditional epidemiological function of the register of consultations is to calculate the incidence of symptoms like injury, stomachache, diarrhea, or acute respiratory infection, but also of tentative diagnoses that can be regarded as more or less satisfactory surrogates for a correct diagnosis of certain ailments, especially of frequent infectious diseases like measles and dengue fever. This is obviously conditional on the training of the health worker whom the patient is consulting.

Registers in policlinics, hospitals, specialized malaria or tuberculosis stations, and the like concern mainly the units "outpatient consultation," "hospitalization," "discharge," and various "laboratory tests." Rare are the hospitals in developing countries that have a so-called *Central Register* whose unit is a "patient." Most of the more elaborate incidence and prevalence statistics which appear, for example, in health yearbooks originate from hospitals where diagnoses are on the whole more precise. Many of the cases treated only in CHCs are lost in this way, though. The old problem of linking records in primary, secondary, and eventually tertiary health care usually does not get sufficient attention.

In particular, the case reports based on consultations and treatments in hospitals may still be quite incomplete. For instance, in Vietnam the cancer incidences reported within that system and published in the Health Yearbook are very much lower than the figures obtained from the Central Cancer Registry that has been operating for some time in a few regions (Krickeberg 1999).

As said in the preceding section, calculating and filing of the corresponding reports ought to be integrated with the data entry into the registers. The details depend of course on whether the register is on paper or on a computer disk. In any case, the report should display the epidemiological situation reported also immediately to the health worker himself in a vivid fashion in the spirit of the "epidemiology at the basis." This brings us back to the "local mastership" over information alluded to above which is one of the main roles of a good HIS, in contrast to the old bureaucratic idea that such a system exists mainly to produce indicators for higher-up administrations. This local mastership must not be taken away from the basic health institutions. It allows them to plan in particular preventive measures on their level. The often advocated "feedback" is useful but after all a poor substitute for local use of information.

Incidences calculated and reported may sometimes be *corrected* later, that is, replaced by *estimates* that will generally be closer to the true incidences and should therefore be used in health statistics and for health planning (Krickeberg 1994). A very rough procedure actually in use is the multiplication of every indicator obtained from the records by a constant "correcting factor" that had been estimated beforehand. An interesting application concerns maternal mortality (Ministerio de Salud Pública y Asistencia Social 2002) where the factor 1.58 is being used. Slightly more elaborate but still elementary estimations could use simple extrapolations or Bayesian methods (Krickeberg 2007).

A HIS fulfills other functions in addition to normal health statistics. Let us mention some of them here. When well organized, the register of consultations in a CHC can serve the *epidemiological surveillance* of infectious diseases as well. Moreover, by defining the values of the variable "symptoms" or "syndrome" appropriately so as to identify the consultations that belong to a program like CDD, ARI, malaria, or tuberculosis, a separate register of consultations for these programs becomes superfluous at the level of the commune without making the general register of consultations any more complicated. The latter will furnish all the incidences required.

The register of consultations of a CHC may also allow some rough but illuminating *experimental epidemiology*. In some CHCs this register contains an additional variable "outcome" which can take the value "cured" or "deceased" among others. Strictly speaking, this variable concerns the larger unit "case" and not a "consultation," but in the absence of a register of cases, it can easily be recorded in a paper-based register of consultations by going back to the first consultation for a case, provided that the outcome becomes known not too late, for example, for acute diseases. When using computers, recording an outcome which may occur much later, or even deriving a register of cases from that of consultations, is fairly easy. The variables "treatment" and "outcome" now permit both the local health workers and health administrators, in particular those of vertical programs, to monitor standard treatments. To this end, they can peruse the complete register during a certain period or, usually more efficiently, employ *sampling from records*, that is, random sampling from the units of a register, an unfortunately fairly neglected method although classical in hospitals of developed countries.

Sampling from records in CHCs also enables health authorities to estimate some of the epidemiological characteristics of the preliminary clinical diagnoses described in Sect. 23.2.3. If the true diagnosis will *eventually* be known for most cases as it happens with malaria, it suffices to retrieve this diagnosis for the subjects included in the sample from the relevant records of district or provincial malaria stations, hospitals, etc. In the opposite case, for example, for dysentery, a laboratory-based diagnosis for the sampled subjects needs to be done.

In addition to the register of consultations, CHCs frequently have a register for the program MCH whose underlying unit is not always well defined. In principle, it should serve the needs of EPI, too, making an EPI register superfluous.

Some CHCs have also set up a register of all *families* or *households* of the commune which can furnish certain epidemiological insights, for example, on infectious diseases or the role of socioeconomic factors in health.

Finally, HISs can be used for epidemiological studies. It is in this context that the "golden rule" formulated in the preceding section plays its main role. If we want to study the influence of a risk factor on an outcome variable, it is necessary to record their values in the *same* register for every unit of the target population. Let us look again at MCH. It is customary to record data on pregnancies in one register and those on the health of the newborn up to a certain age in a different one. This makes it impossible to use the registers, for example, for an investigation of the influence of prenatal care on the development of the child. A natural unit for an MCH register would be "a child from the moment of its conception until the age of 5 years, say." Registers of this kind have been designed by ingenious district health officials in Laos. We will not pursue the issue of epidemiological studies based on registers further; cf. Krickeberg (2007).

23.2.5 Epidemiological Sample Surveys

Let us start again with a truism that does not have much weight in practice either: sample surveys should be done and *only* be done if the results are not readily available from registers. For example, in developing countries many cases of measles do not lead to a contact of the patient with the health system and are therefore registered nowhere, not even with a tentative diagnosis. This is the so-called iceberg phenomenon. In such a situation, to get an idea of measles incidence, sample surveys are required. Another example of a different nature: the causes of death of children under 5. These causes are often not indicated in any register because it is impossible to perform routinely the necessary diagnoses.

We will also discuss situations where sample surveys are *not* required but often still being done.

In the present section, we will restrict ourselves to sample surveys whose purpose is to estimate indicators that depend on a *single* variable, for example, the incidence or prevalence of a disease, or mortality. Epidemiological studies to investigate the *association* between several variables like those between a risk factor and a disease, or between a treatment and the outcome, will be taken up in the last section although the borderline is not everywhere well traced; cross sectional studies, in particular, usually have outwardly the form of a classical sample survey.

The most frequent *units* of a sample survey, that is, elements of the "target population," are a person, a family (household), or a community, for example, a hamlet or a commune, but a unit like a CHC may also occur if one wishes to evaluate the performance of such centers.

It is customary that for every new vertical program, a sample survey is launched in order to estimate the so-called baseline indicators, and often this survey is repeated later to monitor changes that may reflect the impact of the program. Such indicators are, for instance, incidences or prevalences of the diseases; case fatalities; infant, child, or maternal mortalities; indicators describing the nutritional status of children; vaccination coverages; or sometimes plain demographic indicators.

Surveys outside of any vertical program that include an epidemiological component have mostly been organized in the form of a large international project. An early model was the "World Fertility Survey" of the International Statistical Institute (Cleland and Scott 1987) that had been conducted between 1974 and 1982 in 42 developing and a few developed countries. Its methods including software were subsequently incorporated into the baseline surveys of some family planning programs. UNFPA (United Nations Fund for Population Activities) organized "Demographic Health Surveys" (DHSs) in many countries, often repeatedly, and these were later also conducted by private firms. Household surveys done by the United Nations have been a similar global venture (United Nations 1984).

Unfortunately, in a given developing country, these multitudinous surveys have practically never been coordinated with each other regarding their purpose, choice and definition of variables and indicators, methodology, and availability of the results. Therefore they cannot be compared with each other and exploited only with difficulty or not at all. Often, they duplicate existing knowledge, or the results are buried in the files of the relevant program office, or both. By contrast, the World Fertility Survey is an early model of good quality and availability of its results, usually published locally.

Moreover, the definition of variables and indicators employed by most sample surveys is often not compatible with that of existing HISs. The fundamental question to which extent an information system could have furnished the same result is rarely asked. For example, in a country like Vietnam that, although economically poor, provides well-structured primary health-care, most cases of severe diarrhea lead to a consultation in a basic health care institution where the syndrome is recorded. Hence, a large if not the largest part of the CDD surveys that have been conducted were actually superfluous and could have been replaced by "register-based" studies, that is, sampling from records. The opportunity to compare indicators obtained from a survey with those extracted from the HIS is mostly lost although recently there have been notable exceptions, for example, the survey from Guatemala on maternal mortality already quoted (Ministerio de Salud Pública y Asistencia Social 2002).

The sample surveys we are discussing are simple cross-sectional studies that concern in principle only the state of affairs at the moment the data are recorded. If, for instance, we are interested in the incidence of an acute disease or the mortality by it as it happens in the framework of EPI, the diagnosis for a case or the cause of a death has to be recorded for a *period of the past* and can therefore only be determined by asking questions about the past, that is, by an *a posteriori diagnosis* or an *verbal autopsy*, respectively. These questions are addressed to persons in the surroundings of the patient, mostly members of his family.

Sometimes, this may not be very reliable. The average errors committed depend of course on the particular context and can be estimated by elementary studies. In any case, the survey should be confronted with the HIS if there is one. In particular, the questionnaire of the survey ought to contain a question about whether the patient had consulted his or her CHC or another health facility or not. Comparing the answers with the entries in the relevant register of consultations would then tell us a lot about the functioning of the HIS and the health system itself.

There is a related problem in the case of indicators defined with respect to *cohorts* that start at birth like infant mortality, child mortality, or vaccination coverage. In their definition, a cohort of children is to be followed from birth on to the age of 1, 2, or 5 years, respectively. They can be estimated directly by a register-based study. By contrast, a sample survey is done at a fixed moment. It either estimates a different indicator that concerns the state of a group of children at this moment as it happens with the usual EPI-coverage surveys or needs to resort to questions about the past. If the latter strategy is adopted as it is, for instance, being done within the Demographic Health Surveys (DHSs) when estimating child mortality, care must be taken to analyze possible bias resulting from censuring the 5-year period of the past within which a child was to be followed.

To look at the interplay between sample surveys and register-based studies is also most illuminating when dealing with chronic diseases, especially non-infectious ones. Sample surveys done in Vietnam by an association of cardiologists have shown a recent dramatic increase of the prevalence of hypertension, most of it hitherto hidden. Linking such surveys to the HIS could have provided much useful knowledge.

Sample surveys, like HISs, can play an educative and training role, too, as it happened with the World Fertility Survey that largely relied on local staff. The organizers of a survey who often come from an international organization should always take great care to explain to the local health personnel involved its rationale including both the objectives and the methodology. They do not do it every time, partly because not all of them understand the underlying principles of the methods. They also tend to apply ready-made recipes that do not take into account the local administrative structures. Worse, existing know-how is often not exploited; in some countries certain local personnel is in fact better qualified than the foreign "experts" who direct a survey.

For instance, the so-called 30-cluster-sampling plan has been widely used in developing countries (Henderson and Sundaresan 1982). Its first stage is a Madow sampling design (Cochran 1977, p. 265). Its second stage is tailor-made for countries that may not have lists of all households nor suitable maps of hamlets or villages. It has been very useful in many situations because it provides a standard method. Although it is not always self-weighting, it has been treated as if it were which implies a simple form of the estimators. However, it has also been much applied in contexts where other sampling plans would have been easier and cheaper to implement and more efficient. This author has seen the Madow design used by a foreign CDD specialist to select a sample of 30 administrative units from a list of only 32 or 33 such units. The local staff would simply have taken all of these units! Nowadays, suitable maps of houses can be cheaply and rapidly drawn by the geographical positioning system even for very remote and poor villages.

Another example: in a developing country that operates a high-level research institute for demography and family planning, a baseline survey for a reproductive health program was conducted by a person from abroad using the above-mentioned World Fertility Survey software although surveys providing a large part of the results had already been done by that institute and although the institute could very well have done the entire survey itself in a much cheaper and more imaginative manner, profiting at the same time from the training and insights (and money) that go with it.

As said above, baseline surveys of vertical programs and other surveys are sometimes repeated to monitor changes. To this end, samples are taken repeatedly in an independent fashion. There always exists a splendid opportunity to retain part of the sample and to use it for *longitudinal studies* that are most informative, for example, household-based ones, but this opportunity is usually missed as it happened in the family planning survey just mentioned.

The need for longitudinal studies was also one of the motivations for creating the concept of a sentinel network. This is a hybrid between a HIS and a sample survey. Since basic health facilities are all too often ill-equipped for making reliable diagnoses, do not have sufficient staff to keep registers and file reports, and cannot rely on good postal or other services to transmit information, the idea arose to select a fixed sample of such institutions, for example, CHCs. Their capabilities were to be enhanced substantially by furnishing some laboratory equipment, paying additional trained staff, and ensuring better communications. These fixed "sentinels" were to monitor demographic and epidemiological indicators over time including the impact of interventions. Sentinel networks may, however, give biased information because the measures taken to upgrade the selected sites alter the health situation there. More people might be attracted to use these facilities, better treatment can influence the dynamics of the transmission of infectious diseases, and better health education will reduce the incidence of many ailments including injuries.

Originally, sentinel networks were discussed mainly in the context of single vertical programs but occasionally also for larger components of primary health care. There was no general scheme and no coordination. More recently, Demographic Surveillance Systems (DSSs) have been installed in various "field sites," that is, geographically defined populations, in order to monitor their health and demographic status. Let us look at two examples. The first one is the DSS established in the early 1980s by an already existing institution, namely, the Nouna Health Research Center in the northwest of Burkina Faso. It covers about 55,000 people (Yé et al. 2002). The second one is a DSS created during the years 1997–1999 in the context of a Vietnamese-Swedish collaborate research project in the district of Ba

Vi of the province of Ha Tay in the North of Vietnam which counts around 235,000 inhabitants. It is combined with a field laboratory for epidemiological research. First results concern among others mortality, death reporting, theory and practice of verbal autopsies, knowledge of people about tuberculosis, and cardiovascular diseases (Ngyuyen and Diwan 2003).

Starting in 1998, an "International Network of field sites with continuous Demographic Evaluation of Populations and Their Health in developing countries" (INDEPTH) has been built up in order to link existing field sites and in particular to strengthen them and to enhance their visibility and use (INDEPTH 2002). In 2010, it comprised 38 field sites from 18 countries, among them the DSSs at Nouna and Ba Vi.

23.2.6 Epidemiological Studies

We will now take a quick look at how epidemiological studies, other than register-based studies or straight sample surveys, can satisfy the specific epidemiological needs of developing countries. We will thus be dealing with studies of the kind treated in chapters ►Descriptive Studies, ►Cohort Studies, ►Case-Control Studies, ►Modern Epidemiological Study Designs, ►Intervention Trials, and ►Design and Planning of Epidemiological Studies of this handbook.

The practical problems that a research team might face are summarized in Wawer (2001). Let us look at an example that illustrates one of the worst obstacles, namely, loss to follow-up. This study, the Bloemfontein vitamin A trial (Chikobvu and Joubert 2003), had in fact a dual purpose: firstly, to investigate the effect of vitamin A on the transmission of HIV from mother to child and secondly, to assess the extent of loss to follow-up in the presence of AIDS for the planning and analysis of other cohort studies on HIV/AIDS including clinical trials. A total of 303 HIV-positive pregnant women were recruited with 152 randomly allotted to the vitamin A group and 151 to the placebo group, to be followed until their baby was 18 months old. The tracing procedure of these women was precisely defined, the observed loss to follow-up was analyzed in detail, and the kind of bias that may have resulted was discussed. This trial and ten other cohort studies on HIV/AIDS quoted for comparison had losses to follow-up in the order of 10–48%. One of the main conclusions was that published studies must adequately discuss the tracing methods used.

Many sample surveys, especially large ones like the World Fertility Surveys or the Demographic Health Surveys, collect data both on risk factors, for example, social ones, and on outcome variables, for instance, morbidity. They can therefore be exploited in the form of cross-sectional studies on the influence of these factors on such outcomes. A good example is the study from Guatemala already quoted (Ministerio de Salud Pública y Asistencia Social 2002). This involves the shortcomings of cross-sectional studies in general which arise from the fact that many variables concern the past of a cohort and not the moment of the survey; we have sketched these problems in the preceding section.

Sample surveys and cross-sectional studies account for the majority of epidemiological investigations in developing countries. Pure case-control or cohort studies, to be conducted in accordance with a rigid protocol fixed in advance, are still relatively rare. In some developing countries, their number is increasing, though; they are often conducted by a mixed team of local and foreign epidemiologists. It would be futile to enter into details here since the methodology is the same as in developed countries.

The choice of the subject of a study is of course often dictated by local problems that were already mentioned in Sect. 23.2.3. We note that developing countries have been a fertile ground for research in *genetic* epidemiology. For instance, studies on thalassemia have a long history in Southeast Asia. Genetic susceptibility to infectious diseases has been studied intensively. An early example was the evidence for genetic susceptibility to leprosy obtained by a linkage analysis in a closed population, namely, on a Caribbean island (Abel et al. 1989). In the realm of parasitic diseases, the role of genetic factors in resistance to malaria had already been established in 1982 (Mims 1982). More recent work done in Sudan concerns schistosomiasis (Dessein et al. 1999) and visceral leishmaniasis (kala-azar) (Bucheton et al. 2003).

There have also been *clinical trials*, centering in particular around HIV, for example, on the prevention of mother-to-child transmission.

Much epidemiological work in developing countries has the form of *ecological* studies that are treated in chapter ▶Descriptive Studies of this handbook. A good example is the large descriptive ecological study in China (Chen et al. 1991). A community-based intervention trial, the Gambia Hepatitis Study, is mentioned in chapter ▶Intervention Trials of this handbooks, and a few other intervention studies of various sorts are quoted in Wawer (2001). Community-based intervention studies are often a useful and feasible type of investigation given the needs and conditions of developing countries, especially in the context of infectious diseases. They suffer from the same basic difficulties as in developed countries, though. Since there are usually only few communities involved, it is very difficult to exclude confounding factors when evaluating such a study by comparing communities "with intervention" with other, control, communities where no intervention had taken place.

The most typical, illuminating, and useful type of study looks different in practice. It is more informal, and *prospective*, usually lasting for many years or even decades. Often it starts small and expands later regarding objectives, study type, and basic population. It may concern anything connected with health: general or reproductive health of a community as time proceeds; nutrition; etiological factors, especially from the environment or social ones; results of person-based or community-based interventions; and often several of these together. The diseases involved are those already mentioned in the present chapter but rarely non-infectious ones.

An excellent summary of 34 such surveys plus a detailed presentation of 12 among them was published in 1997 (Das Gupta et al. 1997). It starts with pioneer studies done in China in the early thirties, it is organized by continents: Asia, Latin America, and Africa, and it provides fascinating reading. A good example is the malaria study in the Garki district in Nigeria, a savannah region. It took place from 1969 to 1976, involved 22 villages with a total population of 7,423, and has also been the subject of a separate monograph (Molineaux and Gramiccia 1980). The topics studied were manifold: demographic and epidemiological, in particular intervention evaluation. The epidemiology of malaria included all aspects: immunology, serology, parasitology, entomology, clinical manifestations, and the influence of various factors like weather and nutrition. The interventions consisted in pesticide spraying and mass prophylaxis. Let us quote some of the many results. No significant difference in mortality between mass-treated and untreated villages was found except for infant mortality. Malaria antibodies existed in people in sprayed as well as in unsprayed environments, and their level was uncorrelated with that of parasitemia except in infants. The intervention reduced, but did not eliminate, the prevalence of the plasmodia involved. Malaria transmission and vector capacity, which had existed on a high level, persisted.

23.2.7 Conclusions

Summarizing finally the respective roles of the three sources of epidemiological knowledge in developing countries, we may say that, roughly, HISs have their roots in the daily work of the health workers, especially in their clinical activities; sample surveys are mostly tied to particular projects or programs and concern mainly the state of affairs at a particular moment; analytical and experimental epidemiological studies provide the deeper knowledge that is indispensable for all useful planning of health strategies.

23.3 The Example India (by A. Kar, A.K. Chakraborty)

23.3.1 Health Information System for Public Health Interventions

23.3.1.1 Organization of Government Health Services

The historical, cultural, economic, and geographical contexts of a developing country like India influence the development of its health system and the ability of this system to accurately capture data and use them for health interventions. The health system in India is a little over 60 years old. In 1947, when the country achieved independence, health infrastructure was practically non-existent and infectious diseases periodically swept through the population, traumatized by food deprivation. The Report of Health Survey and Development Committee (1946) stated that the death rate in the year 1937 was 22.4 per 100,000/year while expectancy of life at birth was 27 years (Table 23.1). Mortality records of British India estimated a maternal mortality rate of 20/1,000 live births. Table 23.2 presents a comparison of mortality in infants and children in British India as compared to the United Kingdom and Wales. Cholera, smallpox, and plague accounted for 2.4%, 1.1%, and 0.5% of all deaths in the year 1932. Tuberculosis mortality rates in cities

Table 23.1 Health indicators India

Indicator	1937[a]	1951[b]	1981[b]	2000[b]
Demographic indicators				

LE at birth	26.9(M)
	26.5(F)
		36.7	54	64.6

Crude birth rate		40.8	33.9 (SRS)	26.1 (1999, SRS)
Crude death rate	22.4/1,000	25	12.5 (SRS)	8.7 (1999, SRS)
IMR	162/1,000 live births	146	110	70
Epidemiological shifts				
Malaria (cases in million)	100	75	2.7	2.2
Leprosy cases per 10,000 population		38.1	57.3	3.74
Smallpox	69,474 (1.1%)	>44,887	Eradicated	
Guinea worm (no. of cases)			>39,792	Eradicated
Polio			29,709	265
Infrastructure				
SC/PHC/CHC		725	57,363	163,181
Dispensaries & hospitals (all)		9,209	23,555	43,322 (1995–1996, CBHI)
Beds (private and public)		111,198	569,495	870,161 (1995–1996, CBHI)
Doctors (allopathy)		61,800	268,700	503,900 (1995–1996, MCI)
Nursing personnel		18,054	143,887	737,000 (1999, INC)

SRS Sample Registration System, *CBHI* Central Bureau of Health Intelligence, *MCI* Medical Council of India, *INC* Indian Nursing Council, *LE* life expectancy, *IMR* infant mortality rate, *SC* subcenter, *PHC* primary health center, *CHC* community health center

[a] Report of Health Survey and Development Committee (1946). In Compendium of Recommendations of Various Committees on Health and Development 1943–1975. Central Bureau of Health Intelligence, Directorate General of Health Services, Ministry of Health and Family Welfare, Government of India. New Delhi

[b] Planning Commission Government of India (2002)

ranged between 200 and 450 per 100,000 population. Only 2% of the population had access to potable water, while sanitary facilities were available for 4.5% of the populace. Trained health personnel and infrastructure were practically non-existent. This was more pronounced in the rural population, where, for example, one institution was available for the 105,626 inhabitants of 224 villages (Report of Health Survey and Development Committee 1946). From 1948, the evolution of the Indian health system occurred, wherein development of the government health

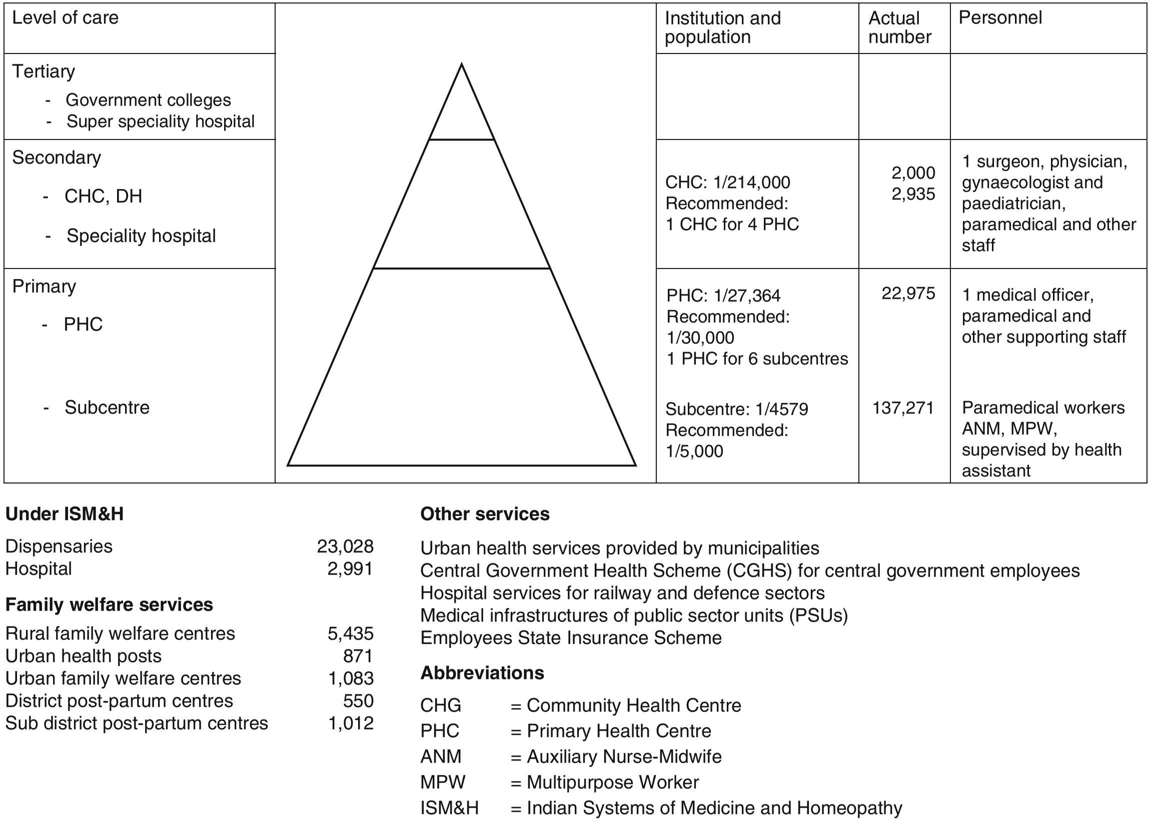
Table 23.2 Deaths at specific age-periods shown as percentages of the total deaths at all ages[a]

	Under 1 year	1–5 years	5–10 years	Total under 10 years
British India (1935–1939)	24.3	18.7	5.5	48.5
England and Wales (1938)	6.8	2.1	1.1	10.0

[a] Report of Health Survey and Development Committee (1946). In Compendium of Recommendations of Various Committees on Health and Development 1943–1975. Central Bureau of Health Intelligence, Directorate General of Health Services, Ministry of Health and Family Welfare, Government of India. New Delhi

services was matched by a parallel proliferation of the private heath sector. Between 1948 and now, health of the population has improved as demonstrated by significant improvements in a number of health indicators (Tables 23.1 and 23.2). Smallpox and guinea worm have been eradicated, and polio is on the verge of extinction (National Polio Surveillance Project 2011). There has been a substantial drop in the total fertility rate and infant mortality rate while life expectancy has increased significantly (Planning Commission Government of India 2002).

The government health system has the mandate of developing and maintaining the health information system and for using this data for prevention and control of diseases, for health promotion activities, for decision making on provision of basic health services at minimal or no cost, and for outbreak investigation and control. The government health system has a definite organization through which health information is collected and transmitted (Fig. 23.1). The first contact between the population and the government health system occurs through the most peripheral health institution termed the subcenter. As mentioned earlier in the chapter, in a developing country like India, the subcenter is manned by paramedical staff, one auxiliary nurse-midwife (ANM) and one multipurpose worker (MPW). These health workers, supervised by a health assistant, have the responsibility of providing basic health-care services to a population of 5,000 residents in about four to six villages. The subcenter level activities include treatment for minor ailments and injuries, collection of blood smears from all fever (suspected malaria) cases, and referring patients and suspects to the Primary Health Center (PHC). Deliveries and antenatal care services are also done from the subcenters. The PHC is the referral unit for six subcenters and represents the next level of the health-care system. All PHCs have at least one qualified medical officer in charge. A majority of PHCs have four to six inpatient beds. One PHC is available for every 30,000 population. The PHC is the hub for compiling health data collected at the subcenters. The Community Health Center (CHC, also termed the First Referral Unit) is a 30-bed hospital that is the First Referral Unit for four PHCs. One CHC is available for every 200,000 population. The secondary health-care infrastructure consists of the district hospitals and other specialty hospitals, while the tertiary health-care infrastructure consists of government hospitals. The structure of the health system is described in Fig. 23.1. Starting from the subcenter level, morbidity and health data are collated and forwarded through district health officers to the Directorate of Health Services of each state. The national level data are collated from 28 states and 7 union



Level of care		Institution and population	Actual number	Personnel
Tertiary - Government colleges - Super speciality hospital				
Secondary - CHC, DH - Speciality hospital		CHC: 1/214,000 Recommended: 1 CHC for 4 PHC	2,000 2,935	1 surgeon, physician, gynaecologist and paediatrician, paramedical and other staff
Primary - PHC		PHC: 1/27,364 Recommended: 1/30,000 1 PHC for 6 subcentres	22,975	1 medical officer, paramedical and other supporting staff
- Subcentre		Subcentre: 1/4579 Recommended: 1/5,000	137,271	Paramedical workers ANM, MPW, supervised by health assistant

Under ISM&H

Dispensaries	23,028
Hospital	2,991

Family welfare services

Rural family welfare centres	5,435
Urban health posts	871
Urban family welfare centres	1,083
District post-partum centres	550
Sub district post-partum centres	1,012

Other services

Urban health services provided by municipalities
Central Government Health Scheme (CGHS) for central government employees
Hospital services for railway and defence sectors
Medical infrastructures of public sector units (PSUs)
Employees State Insurance Scheme

Abbreviations

CHG	= Community Health Centre
PHC	= Primary Health Centre
ANM	= Auxiliary Nurse-Midwife
MPW	= Multipurpose Worker
ISM&H	= Indian Systems of Medicine and Homeopathy

Fig. 23.1 Schematic representation of the public health infrastructure

territories of the country. The magnitude of the health information system can be conceived from the fact that government health services extend beyond urban areas to populations living in the nearly 668,000 villages in the country. The Rural Health Statistics 2009 lists 145,894 subcenters, 23,391 primary health centers (PHCs), and 4,510 community health centers (CHCs) functioning in the country (National Rural Health Statistics 2009).

23.3.1.2 Private Health Services and Traditional Medical Providers

The challenge to the HMIS in India comes from the fact that in addition to the government health services, India has a large private health sector. This private health sector consists of general and specialist medical practitioners, either in independent practice or working from hospitals. The exact number of practicing private practitioners in the country is unknown. Estimates of private practitioners are usually made from the number of students graduating from medical colleges each year. The majority of graduating physicians work in the private sector. The role of the private sector is to provide curative services for a fee. The infrastructure and quality of services present an outstanding range. The majority of the private sector institutions are single doctor dispensaries with very little infrastructure or paramedical support. Specialty care accounts for only 1–2% of the total number of institutions, while corporate hospitals (i.e., hospitals funded by large business organizations) constitute less than 1% of this infrastructure (Report of the Steering Committee on Health for the Xth Five Year Plan (2002–2007), 2002). The latter cater to people from the higher socioeconomic strata and offer their services to medical tourists from other countries. There are no mechanisms for collecting, integrating, and utilizing information on disease, infrastructure, or manpower from the private sector. In addition to the private health sector, it is estimated that there are more than 7,000 voluntary agencies (NGO, non-governmental organization) are involved in health related activities, mostly in the rural areas Report of the Steering Committee on Health for the Xth Five Year Plan (2002–2007) (2002).

Data of the National Sample Survey Organization show that there are significant interstate differences in the distribution of private sector hospitals and beds. Private sector hospitals are present in more prosperous districts and states, while in the poorer districts where health infrastructure is most needed, health services are provided by the public health sector (National Sample Survey Organisation 1998). The vulnerability of the HMIS is increased by the presence of multiple systems of medical practice in India. In addition to allopathic practitioners, licensed practitioners of traditional Indian systems of medicine (ayurveda, unani, siddha), Tibetan medicine and homeopathy are widely used by people in all parts of the country. A new Department of Ayurveda, Yoga, Unani, Siddha and Homeopathy (AYUSH) has been created (Anonymous 2008) with the objective of promotion and propagation of these systems of medicine in India and globally. There is a vast network of governmental institutions practicing and promoting the Indian Systems of Medicine and Homeopathy (ISM&H). There are 3,004 hospitals with 60,666 beds and over 23,000 dispensaries providing primary health care (Report of the Steering Committee on Health for the Xth Five Year Plan (2002–2007), 2002).

Over 16,000 practitioners from 405 colleges qualify every year. There is no estimate of the utilization of this sector, although ISM&H practitioners are extremely popular throughout the country. Inclusion of morbidity data from these practitioners into a disease database would be greatly limited since the rationale for diagnosis and disease classification is entirely different from the allopathic system of medicine.

23.3.1.3 Health and Morbidity Data: Disease Surveillance

Data on reproductive and child health, vector-borne diseases, tuberculosis, leprosy, blindness control, iodine deficiency disorders as well as health infrastructure, district level health plans for different states, community monitoring, etc., can be obtained through the portal of the National Rural Health Mission. Some selected sources of data are listed in Table 23.3. Data collection formats from subcenter to national level can be viewed at http://nrhm-mis.nic.in/Downloads_HMIS.aspx. The records reflect the health priorities of the country: maternal health (primarily deliveries and outcome), maternal and child immunizations, endemic communicable diseases, health infrastructure, and health manpower (the latter necessary for delivering services through the 1.2 billion population). Information is also collected on deaths and probable causes of death. Morbidity statistics on major infectious diseases are collected by the "vertical" disease control programs (described in part A) from both rural and urban health institutions. Again, the magnitude of the HMIS becomes apparent from the example of the tuberculosis control program in the state of Maharashtra. Data are collected from 29 district tuberculosis centers and 1995 peripheral health institutions, and 7 tuberculosis hospitals and sanatoria (Health

Table 23.3 Website links for HMIS for various health and disease control programs

National Rural Health Mission	http://mohfw.nic.in/NRHM.htm
Reproductive and child health	http://mohfw.nic.in/NRHM/RCH/Index.htm
National Vector Borne Disease Control Program	http://nvbdcp.gov.in/
Revised National Tuberculosis Control Program	http://www.tbcindia.org/
National Iodine Deficiency Disorders Control Program	http://mohfw.nic.in/NRHM/niddcp.htm
National Leprosy Eradication Program	http://nlep.nic.in/
"Pre conception and Pre Natal Diagnostic Techniques (Prohibition of Sex selection)" Act -1994	http://rajswasthya.nic.in/PCPNDT.htm
Bulletin on Rural Health Statistics	http://mohfw.nic.in/NRHM/BULLETIN%20ON.htm
Community based monitoring of health services	http://www.nrhmcommunityaction.org/
Health statistics, causes of death, etc.	http://nrhm-mis.nic.in/Publications.aspx
Pandemic influenza	http://mohfw-h1n1.nic.in/index.html
Sample Registration System	http://censusindia.gov.in/Vital_Statistics/SRS/Sample_Registration_System.aspx

Status Report Maharashtra 2009). Each center is responsible for recording and reporting clinical data on cases, treatment outcome, case holding, and management. Each parameter is strictly defined and follows the guidelines from the global program. It is pertinent to recall that the HMIS is based on health data from the government system, since there is no mechanism to incorporate data from private practitioners, used by the large majority of the population.

The Integrated Disease Surveillance Project is a surveillance program for early warning of impending outbreaks. Although the private sector provides over 75% of curative care for common illnesses, the disease surveillance system does not have representation in the private health sector. Usually, diseases which are considered to be serious menaces to public health, like measles, whooping cough, and diphtheria, are included in the list of notifiable diseases. In India, there is a conspicuous lack of uniformity in the lists of diseases which are notifiable among the 28 states and 7 union territories of the country. Cholera, yellow fever, and plague which are internationally quarantinable diseases are notifiable throughout the country as required under international health regulation. Other than these three diseases, most of the important infectious diseases are not uniformly notifiable throughout the country.

23.3.1.4 Sources of Population Data

Denominator data for epidemiological studies can be obtained from a number of sources, such as the vital registration system, Sample Registration System, and from the census data. Although vital registration is required under law, there is significant under-registration especially in rural areas. Monitoring of vital events is also difficult considering the magnitude of the task. In Maharashtra, the second largest state in the country, for example, vital statistics data are obtained from 40,448 villages, 230 municipal councils, 15 corporations, 7 cantonment boards, and 4 ordnance factories. It is estimated that in Maharashtra, about 80% of births and about 62% of deaths are registered under the civil registration system (Health Status Report Maharashtra 2003). Computerization of vital events is present in only a handful of municipalities in the country. Vital registration is likely to be better in urban areas, since death certificate is compulsory for obtaining permission for cremation or burial. This rule is rendered ineffective in rural areas where regulated cremation areas are often not used (Health Status Report Maharashtra 2003).

A more reliable estimate of vital rates is obtained from the Sample Registration System (SRS). The objective of the SRS is to provide reliable annual birth and death rates at the state and national level for urban and rural areas. The SRS is a dual system of registration involving continuous enumeration of births/deaths in sample villages/urban blocks and matching these data with that of a 6-monthly retrospective survey. The data obtained through these two sources are verified by matching. Unmatched data are reverified. This method not only results in unduplicated reporting of events but also has an inbuilt system of quantifying the disparity between the two data sets. The details of sample design and sample size are given on the SRS website (Sample Registration System (SRS) Bulletin 2001) (http://censusindia.gov.in/Vital_Statistics/SRS/Sample_Registration_System.aspx).

The census forms one of the more authentic sources of demographic data in India. The first population census was initiated between 1865 and 1872 and has subsequently been undertaken uninterruptedly once every 10 years. Post enumeration surveys (PES) form an integral part of the census operation, aimed at estimating coverage and content error. Error tables for under-coverage of census houses and population are available through periodic publications from the Office of the Registrar General and Census Commissioner. Content error is estimated from a 10% sample of households enumerated in the census and reverifying certain specific data in the questionnaire. The 15th Census 2011 will lead to the formation of a National Population Register (NPR). The NPR will contain identification data of the population, which is to be linked to a Unique Identity Number. This system, once operational, would be of great relevance to epidemiologists. Currently, since health care is out of pocket, mechanisms for tracking patients, eliminating duplicate data, etc., form a formidable challenge to epidemiologists working in the country.

A series of Demographic Health Surveys (the National Family Health Surveys) have yielded accurate and valuable information. The primary purpose of these National Family Health Surveys was to collect data on key indicators for assisting in policy decisions pertaining to the national population policy. The utility of these surveys is demonstrated by the last survey (NFHS-3) where HIV testing was done for over 100,000 men and women aged between 15 and 49 years. HIV estimates in India were previously being made from sentinel surveillance of infection rate among pregnant women attending antenatal clinics (ANC) and from sentinel sites that were not completely representative of the Indian population due to the convenient nature by which the sites were selected. Using this far from accurate surveillance system, the National AIDS Control Organization and international agencies had estimated the adult prevalence of HIV to be 0.91%. The NFHS-3 survey reported one third of the estimate of the earlier surveillance system at 0.36% (NFHS-3 2007), leading to downward estimates of not only the Indian but also global HIV prevalence estimates.

23.3.1.5 HMIS Data and Their Limitations

Despite the existence of this extensive health collection system, there are two major reasons contributing to the incompleteness of the existing HMIS. As previously mentioned, the first reason is the lack of a mechanism to collect data from the health-care providers in the private sector. Various surveys have shown that people from all socioeconomic strata access the private sector for outpatient services. For inpatient services, 60% of individuals living below the poverty line (26% of the Indian population fall into this category) utilize the public health sector facilities. An equal number of people above the poverty line access the private health sector (National Sample Survey Organisation 1998, 52nd Round 1998). Other surveys have shown that 65% of households go to private hospitals or doctors for treatment when a family member is ill. Only 29% usually approach the public medical sector. Even among poor households, only 34% normally use the public medical sector during illness (NFHS-2 2000). A more recent survey of a nationally representative sample of 109,041 households found that nearly two thirds (65%) of households use the private medical sector for health care (NFHS-3 2007). The second reason contributing to the incompleteness of the data is that unlike rural health services, urban health services do not have the planned and organized primary, secondary, and tertiary services in geographically delineated areas. Usually, private practitioners, municipal dispensaries, and tertiary care institutions cater to an enormous patient population from both the urban and rural areas. There is no population identifier, and since nearly all health-care expenditure is out-of-pocket, there is no estimate of duplications of data. Thus, the enormous population, the mixed health system, and the lack of initiatives to integrate data from the government and private health-care sectors make the task of information management and using this data for planning health interventions an intimidating task.

23.3.2 Epidemiological Research in Shaping Public Health Programs: Tuberculosis Control

In this section, we demonstrate how epidemiological research from major research institutions together with the public HMIS have helped shape disease control programs, using the specific example of tuberculosis. India is listed first among the 22 high-burden countries that account for 80% of new tuberculosis cases in the world (World Health Organization Report 2009). Although the incidence of tuberculosis has decreased, in absolute numbers the country still has the largest number of prevalent cases of tuberculosis globally. Some seminal epidemiological investigations conducted by Indian research institutions led to defining tuberculosis control policies in the country. Tuberculosis is a long-standing, chronic epidemic which has existed for centuries in India, being recorded in the ancient ayurvedic texts. The first national survey aimed at estimating disease burden was undertaken shortly after independence when a national sample survey for tuberculosis was conducted in the country (Indian Council of Medical Research 1959). This survey measured the prevalence of tuberculosis using a sample of 300,000 individuals drawn from 150 villages, 6 cities, and 30 towns (Chadha et al. 2005). The survey showed that tuberculosis was a significant public health problem in the country, which required an increase in treatment facilities and a public policy for control of the disease. The manner by which treatment for tuberculosis patients could be delivered in an affordable manner in a resource poor country like India in the 1950s and 1960s was defined by a number of other key research studies. In 1956, the so-called Madras study demonstrated that ambulatory tuberculosis treatment was effective and that it did not increase the risk of tuberculosis amongst family contacts (Tuberculosis Chemotherapy Centre, Madras 1959). The implications of this study were enormous, since tuberculosis patients did not require admission for the duration of their treatment. In 1963, it was demonstrated that most tuberculosis cases approached treatment facilities due to persistence of symptoms (Bannerji and Andersen 1963). This significant study implied that there was no need for active case finding for tuberculosis. The feasibility of passive case finding using sputum smear microscopy for presence of acid-fast bacilli (AFB) opened the way for establishing relatively low-cost tuberculosis diagnosis and treatment facilities in Primary Health

Centers throughout the country (Baily et al. 1967). Based on these and other studies, the National Tuberculosis Program (NTP) was initiated in 1962, having as its core, case detection through sputum microscopy of chest symptomatics attending government health facilities on their own (supplemented by X-ray when needed) and ambulatory treatment of 12–18 months duration. The NTP was reviewed in the 1990s, and the Revised National Tuberculosis Control Program (RNTCP) having components of the global tuberculosis control initiative was launched from 1992. Key to RNTCP was supervised treatment of patients, directly observed treatment, and short course (DOTS) (World Health Organization 1994). Research conducted in 1958 and in the 1980s in India had shown the impact of treatment supervision on improving cure rates and that intermittent treatment was as efficacious as a daily treatment regimen (Tuberculosis Chemotherapy Centre, Madras 1959; Tuberculosis Research Centre Madras and National Tuberculosis Institute Bangalore 1986). More extensive reviews on tuberculosis control in India are available from some key articles and references quoted therein (Chakraborty 1997; Narayanan et al. 2003; Chadha et al. 2005).

23.3.2.1 Limitations of the Tuberculosis Register and Role of Research

The RNTCP is the example of a vertical program, having its own structure, manpower, precise case definitions, patient categorization guidelines, treatment guidelines, and data-reporting formats. Details can be viewed at http://www.tbcindia.org/documents.asp#. These extensive data serve to guide the governmental tuberculosis control program. As described earlier, tuberculosis case notifications in this disease register is from the public health program only and therefore does not represent the true prevalence and incidence of tuberculosis. The number of cases detected by the government program is expressed as the case detection rate, calculated as the number of annual new smear positive (infectious cases) notifications from the RNTCP to the estimated new smear positive incidence. The estimates of tuberculosis incidence (and prevalence) come from repeated studies conducted in different Indian institutes. India now has a case detection rate of 68% indicating that a large number of patients remain unnotified, probably accessing tuberculosis treatment from private medical practitioners (World Health Organization 2009).

23.3.2.2 Estimation of Burden of Disease Through Surveys

Tuberculosis infection is estimated through tuberculin surveys. Tuberculin test results in an induration which is measured between 48 and 96 hours of tuberculin challenge. The size of the induration is plotted on a histogram. It is possible to identify tuberculin reactors from non-reactors, since the induration measures cluster around two modes, separated by an antimode. For estimation of burden of disease, examination of sputum smears for presence of acid-fast bacilli (AFB) and radiography have been the standard epidemiological survey tools. The specificity of microscopy of sputum is high, ranging between 98% and 99%. However, the sensitivity is relatively poor ranging between 50% and 70% or even lower. Repeated sputum examinations are known to improve the sensitivity of microscopy

(Nair et al. 1976). For example, when two specimens are studied, case detection improves from 58% to 72%. Only 6% additional cases are however detected if 8 specimens are examined. Culture of concentrated bacilli from processed sputum on standard culture medium is more sensitive than microscopy, since culture detects approximately 100 bacilli per milliliter of sample. Although culture is considered to be the gold standard, expense has been a major limitation in the use of sputum culture in epidemiological studies in India. Transportation of samples from remote areas to the laboratory is another major problem that has challenged tuberculosis surveys. Although radiography has also been one of the major tools in epidemiological investigations, X-ray is considered to be a non-specific tool since considerable subjectivity exists in the reading of X-radiographs (Gothi et al. 1974). A list of major surveys and their outcomes have been summarized in Chakraborty (1997). Two major methodologies have been used in prevalence studies. In the "X-ray screening" approach, mass miniature radiography (MMR) of the entire population is used to identify individuals with abnormal chest shadows for sputum tests. In the "symptom-screening" method, chest symptomatics in the community have been identified through questioning individuals for the cardinal signs of tuberculosis (i.e., persistent cough, chest pain, low-grade fever, and hemoptysis). Individuals screened positive by this method are offered sputum examination and chest radiography.

Although prevalence estimates of pulmonary tuberculosis have been computed from different surveys that have been conducted from time to time in the country, the surveys lack in uniformity since the objectives and methodologies have varied (Chakraborty 1997). Moreover, the surveys have been conducted in certain specified and restricted areas of the country covering different populations. Thus, the available information is neither uniform in its content nor representative of the country as a whole. A global exercise has been conducted by Dye et al. (1999) directed toward providing an average estimate for the country. An updated summary of the tuberculosis situation in the country as measured through various surveys is summarized in Chadha et al. (2005). Prevalence estimates from surveys have been used for developing control strategies and for estimating the resources that will be required for the control program. However, such average calculations cannot be used to evaluate changes in the tuberculosis epidemic in the country, for which precise rates, such as those obtained through Annual Risk of Infection (*ARTI*) surveys described below are necessary.

The *ARTI* gives the proportion of the population which will be primarily infected or re-infected with tubercle bacilli in the course of 1 year (Styblo et al. 1969). It is usually expressed as a percentage or cumulative rate. *ARTI* is estimated through the following formula (Cauthen et al. 1988):

$$\widehat{ARTI} = 1 - (1 - P)^{1/A},$$

where A = average age studied and P = prevalence of infection. The method of calculating average age (A) is described in Cauthen et al. (1988). The estimated *ARTI* is actually demonstrated to be the same as the incidence of infection, worked out by repeat testing of the same population under Indian conditions, and represents the only observational evidence ever reported (Chakraborty et al. 1992). *ARTI* studies from different parts of India have demonstrated an infection rate of 1.5–2% per year, but regional differences in infection rates exist within the country (Chadha et al. 2005).

23.3.2.3 Evaluation of Epidemiological Changes Through Time

Assisting the HMIS in understanding tuberculosis in the population, longitudinal studies, conducted in a few areas of the country, has provided an insight into tuberculosis trends. One of the first longitudinal studies, conducted over a period of 23 years, was initiated in a rural population residing in 119 randomly selected villages of Bangalore district. All persons above 5 years of age were X-rayed, and those with radiological abnormality were bacteriologically investigated (Gothi et al. 1979; Chakraborty et al. 1982). The results of this study gave one of the first opportunities to observe the time trend of tuberculosis in a rural community in India. Prevalence and incidence of culture-positive and radiologic cases revealed no change during the period of 12 years for which information was available. The mean age of cases was higher at later surveys. *ARTI* had declined from 1.1% to 0.65% in 23 years. Incidence of smear positive cases had declined for the area from about 65 to 23 per 100,000 in the same period, parallel to the falling *ARTI*. The above data were used to estimate future case rates with the help of a mathematical model which projected that even in 50 years, tuberculosis case rates expressed in terms of prevalence of culture positive cases would come down minimally (Balasangameshwara et al. 1992). It also showed that more energetic and efficient tuberculosis control measures as defined in the model could, however, result in a verifiable change in case rates. The epidemic situation in India is probably on a slow downward curve as indicated by declining mortality and case fatality rates, decline in meningeal and miliary forms of the disease, relatively high prevalence of cases in higher age with a low rate of positive cases in children, higher prevalence of cases in males, especially adult males, and equal prevalence rates across the urban-rural divide. However, even if on a downward trend, the decline could be minimal, as witnessed from high *ARTI* of 1–2% (Chadha et al. 2005 and references cited therein) and an annual decline of around 0–3% reported from the Bangalore rural and Chennai areas (Chakraborty 2004).

23.3.3 Conclusions

The discussion serves to highlight some of the challenges of establishing a HMIS in a developing country. For India, demographic and geographical features of the country determine the magnitude of the HMIS and also challenge timely, complete, and accurate collection of health data. Despite the formidable nature of the task, an extensive governmental reporting system has been put in place in India. These systems use precise definitions, structured data-reporting formats, in many cases developed and shared with global disease control programs, and involve repeated training of staff at all levels of the health system in the method of collecting and reporting data. Despite these efforts, the mixed health system consisting of an unregulated private health-care sector undermines government efforts at collecting health information. For India at least, the HMIS at best reflects disease trends among those utilizing public health resources. In the second half of this section we show the role of epidemiological research in supporting decision making in public health programs through research studies. Thus, the limitation of absence of a disease data capture system from the private sector is supplemented through research that provides data for decision making. We conclude by noting that infrequent and isolated surveys are not the long-term solution to acquiring incidence and prevalence data (Dye et al. 2003). Rather, the ultimate solution would be the same approach currently used in low incidence countries, namely, comprehensive routine disease surveillance. Daunting as the task may appear, it is obvious that strengthening disease surveillance and HMIS systems need to be given a high priority in developing countries like India. A cohesive research thrust is needed to determine the mechanisms by which the existing HMIS can be extended so as to also include health information from the private health sector of the country.

Acknowledgements The authors thank Eliza Dutta for her comments on the manuscript and Dr. N.S. Deodhar, Kavita Modi-Parekh, and Mugdha Potnis-Lele for comments on an earlier version of the manuscript.

References

References to Sect. 23.2

Aaby P (1997) Bandim: an unplanned longitudinal study. In: Das Gupta M, Aaby P, Garenne M, Pison G (eds) Prospective community studies in developing countries. Clarendon, Oxford, pp 276–296

Abel L, Demenais F, Baule MS, Blanc M, Muller A, Raffoux C, Millan J, Bois E, Babron MC, Feingold N (1989) Genetic susceptibility to leprosy on a Caribbean island: linkage analysis with fire markers. Int J Leprosy 57:465–471

Bucheton B, Abel L, Sayda El-Safi, Musa M Kheir, Pavek S, Lemainque A, Dessein AJ (2003) A major susceptibility locus on chromosome 22q12 plays a critical role in the control of kala-azar. Am J Hum Genet 73:1052–1060

Chen JS, Campbell TC, Li JY, Peto R (1991) Diet, life-style and mortality in China. A study of the characteristics of 65 Chinese counties. Oxford University Press, Oxford

Chikobvu P, Joubert G (2003) Follow-up in longitudinal studies in the presence of HIV/AIDS: the Bloemfontein Vitamin trial, a case study. In: Bull International Statistical Institute 54th session proceedings, invited paper meeting 54, ISI, Den Hague

Cleland J, Scott C (1987) The World Fertility survey, an assessment. Oxford University Press, Oxford

Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York

Das Gupta M, Aaby P, Garenne M, Pison G (eds) (1997) Prospective community studies in developing countries. Clarendon, Oxford

Dessein AJ, Hillaire D, Elwali NEMA, Marquet S, Mohamed-Ali Q, Mirghani A, Henri S, Abdelhameed AA, Saeed OK, Magzoub MMA, Abel A (1999) Severe hepatic fibrosis in schistosoma mansoni infection is controlled by a major locus that is closely linked to the interferon-g receptor gene. Am J Hum Genet 65(3):709–721

Henderson RH, Sundaresan T (1982) Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method. Bull World Health Organ 60:253–260

INDEPTH (2002) INDEPTH, health and demography in developing countries, vol 1: population, health and survival at INDEPTH sites. IDRC, Ottawa

Krickeberg K (1989) Intégration des soins de sant'e primaires. Lecture at the Faculty of Medicine of Phnom Penh, January 1989. Séminaire de Statistique Médicale, Université de Paris V (1989–90), pp 45–69

Krickeberg K (1994) Health information in developing countries. In: Simon A. Levin (ed) Frontiers in mathematical biology, Lecture notes in biomathematics, vol 100, Springer, Berlin, Heidelberg, pp 550–568

Krickeberg K (1999) The health information system in Vietnam in 1999. Joint health systems development programme Vietnam-EU, Unpublished report, available from the author: krik@ideenwelt.de

Krickeberg K (2007) Principles of health information systems in developing countries. Health Inf Manag J 36(3):8–20

Lippeveld T, Sauerborn R, Bodart C (eds) (2000) Design and implementation of health information systems. World Health Organization, Geneva

Mims (1982) Innate immunity to parasitic infections. In: Cohen S, Warren K (eds) Immunology of parasitic infections. Blackwell, Palo Alto, pp 3–27

Ministerio de Salud Pública y Asistencia Social (2002) Línea basal de mortalidad materna para el año 2000. Guatemala

Molineaux L, Gramiccia G (1980) Le projet Garki. Organisation Mondiale de la Santé, Genève

Nguyen TKC, Diwan VK (2003) FilaBavi, a demographic surveillance site, an epidemiological field laboratory in Vietnam. Scand J Public Health 31(Suppl. 62): 3–7

United Nations (1984) Handbook of household surveys, Revised edn. Studies in methods, series F, vol 31 United Nations Publications, New York

Wawer MJ (2001) Research collaborations in developing countries. In: Thomas JC, Weber DJ (eds) Epidemiologic methods for the study of infectious diseases. Oxford University Press, Oxford, Chap 21, pp 431–447

Yé V, Sanou A, Gbangou A, Kouyaté B (2002) The Nouna demographic surveillance system, Burkina Faso. In: INDEPTH, health and demography in developing countries, vol 1: population, health and survival at INDEPTH sites. IDRC, Ottawa, pp 221–226

References to Sect. 23.3

Anonymous (2008) Draft Ayush Policy. http://indianmedicine.nic.in/AYUSH-2008%20Draft%20(For%20Website)/Cover%20and%20Contents/0.0.C.%20Index.pdf. Accessed 12 Aug 2010

Baily GVJ, Savic D, Gothi GD, Naidu VB, Nair SS (1967) Potential yield of pulmonary tuberculosis cases by direct microscopy of sputum in a district of South India. Bull World Health Organ 37:875–892

Balasangameshwara VH, Chakraborty AK, Chaudhuri K (1992) A mathematical construct of epidemiological time trend in tuberculosis – fifty year study. Ind J Tuber 39:87–98

Bannerji D, Andersen S (1963) A sociological study of awareness of symptoms amongst persons with pulmonary tuberculosis. Bull World Health Organ 29:665–669

Cauthen GM, Pio A, ten Dam HG (1988) Annual risk of tuberculosis infection. WHO|TB|88.154. World Health Organization, Geneva, pp 1–34

Census of India (2001) Provisional population totals – India, Paper 1 of 2001. Registrar General and Census Commissioner of India, Government of India, New Delhi

Chadha VK, Kumar P, Jagannatha PS, Vaidyanathan PS, Unnikrishnan KP (2005) Average annual risk of tuberculous infection in India. Int J Tuberc Lung Dis 9(1):116–118

Chakraborty AK (1997) Prevalence and incidence of TB infection and disease in India: a comprehensive review. WHO/TB/97.231-World Health Organization, Geneva

Chakraborty AK (2004) Epidemiology of tuberculosis: Current status in India. Indian J Med Res 120:248–276

Chakraborty AK, Singh H, Srikantan K, Rangawsamy KR, Krishnamurthy MS, Stephen JA (1982) Tuberculosis in a rural population of South India: report on five surveys. Ind J Tuber 29:153–167

Chakraborty AK, Chaudhury K, Sreenivas TR, Krishnamurthy MS, Shashidhara AN, Channabasavaiah R (1992) Tuberculosis infection in a rural population of South India: 23 year trend. Tuber Lung Dis 73:213–218

Dye C, Scheele S, Dolin P, Pathania V, Raviglione M (1999) Global burden of tuberculosis. Estimated incidence, prevalence and mortality by country. JAMA 282:677–686

Dye C, Watt CJ, Bleed DM, Williams BG (2003) What is the limit to case detection under the DOTS strategy for tuberculosis control? In: Glimpses. Selected full text articles presented at the 4th World Congress on Tuberculosis. Washington, DC USA June 3–5, 2002. TB Care, India, pp 14–22

Gothi GD, Chakraborty AK, Banerjee GC (1974) Interpretations of photoflurograms of active pulmonary tuberculosis patients found in epidemiological survey and their five year fate. Ind J Tuber 21:90–97

Gothi GD, Chakraborty AK, Nair SS, Ganapathy KT, Banerjee GC (1979) Prevalence of tuberculosis in a South Indian District twelve years after initial survey. Ind J Tuber 26:121–135

Health Status Report Maharashtra (2003) Public Health Department, Government of Maharashtra, Maharashtra

Health Status Report Maharashtra (2009) Public Health Department, State Health Systems Resource Centre, Government of Maharashtra, Maharashtra

Indian Council of Medical Research (1959) Tuberculosis in India: a sample survey, 1955–1958. Special Report series No. 34, ICMR, New Delhi India, pp 1–121

Nair SS, Gothi GD, Naganathan N, Rao KP, Banerjee GC, Rajalakshmi R (1976) Precision of estimates of prevalence of bacteriologically confirmed pulmonary tuberculosis in general population. Ind J Tuber 23:152–159

Narayanan PR, Garg R, Santha T, Kumaran PP (2003) Shifting the focus of tuberculosis research in India. In: Glimpses. Selected full text articles presented at the 4th World Congress on Tuberculosis Washington, DC USA June 3–5, 2002. TB Care India, pp 58–65

National Family Health Survey (NFHS-2), 1998–99 (2000) International Institute for Population Sciences Mumbai India, and ORC. Macro Calverton, Maryland, USA. http://www.nfhsindia.org. Accessed 18 May 2004

National Family Health Survey (NFHS-3), 2005–06 (2007) International Institute for Population Sciences (IIPS) and Macro International. National Family Health Survey (NFHS-3), 2005–06 India: Volume I. Mumbai: IIPS. http://www.nfhsindia.org. Accessed 23 April 2011

National Polio Surveillance Project. http://www.npspindia.org. Accessed 23 April 2011

National Rural Health Statistics (2009). http://www.mohfw.nic.in/NRHM/BULLETIN%20ON.htm. Accessed 12 Aug 2010

National Sample Survey Organisation (1998) Morbidity and treatment of ailments 52nd Round (July 1995–June 1998). Report No. l, 449, Department of Statistics, New Delhi

Planning Commission Government of India (2002): National Human Development Report 2001. http://planningcommission.nic.in/reports/genrep/index.php?repts=nhdcont.htm

Report of Health Survey and Development Committee (1946) In: Compendium of recommendations of various committees on health and development 1943–1975. Central Bureau of Health Intelligence, Directorate General of Health Services, Ministry of Health and Family Welfare, Government of India, New Delhi

Report of the Steering Committee on Health for the Xth Five Year Plan (2002–2007) (2002) Government of India, Planning Commission, New Delhi

Sample Registration System (SRS) Bulletin (2001) Volume 35, No. 1 Registrar General and Census Commissioner. India, Vital Statistics Division, New Delhi

Styblo K, Meijer J, Sutherland I (1969) The transmission of tubercle bacilli, its trend in a human population. TSRU report No 1. Bull Int Union Tuber XLII:1–104

Tuberculosis Chemotherapy Centre, Madras (1959) A concurrent comparison of home and sanatorium treatment of pulmonary tuberculosis in South India. Bull World Health Organ 21:51–144

Tuberculosis Research Centre Madras and National Tuberculosis Institute Bangalore (1986) A controlled clinical trial of 3- and 5-month regimens for the treatment of sputum positive pulmonary tuberculosis in South India. Am Rev Respir Dis 134:27–33

World Health Organization (1994) Global tuberculosis programme. Framework for effective tuberculosis control. Document WHO/TB/94.179. World Health Organization, Geneva

World Health Organization (2009) Global tuberculosis control – epidemiology, strategy, financing Document WHO WHO/HTM/TB/2009.411 World Health Organization, Geneva

Health Services Research

24

Thomas Schäfer, Christian A. Gericke, and Reinhard Busse

Contents

T. Schäfer (✉)
Department of Economics and Information Technology, University of Applied Sciences Gelsenkirchen, Bocholt, Germany

C.A. Gericke
The Wesley Research Institute and University of Queensland, School of Population Health, Brisbane, Australia

R. Busse
Department of Health Care Management, Berlin University of Technology, Berlin, Germany

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_38,

24.1 Introduction

After a brief introduction into the general field of health services research, a large section deals with the specific issues arising when epidemiological or statistical methods are used to study health services. This is followed by sections describing the main fields of investigation which are usually thought of as pertaining to the wider realm of health services research. These are studies of demand, need, utilization, and access to health services which have the interface between the patient and health services in common. The next section describes the importance of financial resources, structure, and organization for the delivery of effective and efficient health care. This is followed by a description of the processes and outcomes of health care, including concepts such as effectiveness and appropriateness of care and their use, for example, in physician profiling or in hospital rankings. In the section on outcomes, special emphasis is put on health status measurement and the evaluation of health systems in international comparisons. Important health economic concepts, such as cost-effectiveness and efficiency, are covered in various sections. This chapter concludes with describing common pitfalls and caveats in interpreting health services research.

24.1.1 Health Services Research Defined

Health services research (HSR) attempts to answer questions about the best medical treatment or preventive course of action, the quality of care provided by a hospital or a physician, the efficient delivery of services to all populations, and their costs. The Institute of Medicine (1994) defines HSR as “A multi-disciplinary field of inquiry, both basic and applied, that examines access to, and the use, costs, quality, delivery, organization, financing, and outcomes of health-care services to produce new knowledge about the structure, processes, and effects of health services for individuals and populations.” The three basic dimensions of care studied are (1) the process of deciding what care to provide, (2) the process of providing care in the best possible manner, and (3) the outcomes that result from care (Scott and Campbell 2002). Many HSR projects study aspects of care that span all three dimensions under the rubric “quality of care” (Brook and Lohr 1985). As Scott and Campbell (2002) pointed out, this frequently used but rarely defined phrase encompasses notions of effectiveness, efficiency, safety, access, and consumer satisfaction and is thus not a very precise title for scientific investigation (Scott and Campbell 2002). HSR challenges the dominant biomedical model in which disease occurs, leading to illness, which is then treated (Black 1997). In contrast to the clinical view focusing on individual patients, it adopts a population perspective and considers other determinants of the use of health care (Black 1997), such as socioeconomic status, local availability or acceptability of health services. HSR thus often challenges medical claims about the value of specific interventions.

HSR cannot be defined as a methodological discipline. It draws upon and uses multiple methodologies and is multidisciplinary in nature. The majority of quantitative research in the field is done using epidemiological methods, and epidemiologists increasingly work in this field of research.

This multidisciplinary approach is seen by many authors as characteristic of this field of investigation, which is reflected in Last's definition of HSR as "The integration of epidemiological, sociological, economic, and other analytical sciences in the study of health services. HSR is usually concerned with relationships between need, demand, supply, use, and outcome of health services. The aim is evaluation, particularly in terms of structure, process, output, and outcome" (Last 2001).

The ultimate goal of HSR, however, is to provide unbiased, scientific evidence to influence health services policy at all levels so as to improve the health of the public (Black 1997).

24.1.2 The Input-Output Model of Health Care

Different models have been proposed for the study of health services. These include operational models, for example, the patient-flow model or the social sciences model. The patient-flow model starts with the assumption of a healthy population, where a patient's way through the different health-care institutions is followed once a disease manifests itself (Bennett 1978). The social sciences model attempts to consider the main social and political influences, causal relationships, and environmental conditions on the process of service delivery in a health-care system. Social experiences, values, priorities, importance of societal resources and structures are the focus of the analysis (Weinermann 1971). A causal or epidemiological model is also possible, which analyses care along known or supposed hypothetical causal biosocial links (de Miguel 1971). The drawback of such a model is its complexity.

For many analyses, simpler models are more adequate. We prefer a model adapted from engineering sciences in which some components of the other models have been integrated – the input-output model (Fig. 24.1) – which takes structure, processes, outputs, and outcomes into account (Schwartz and Busse 2003).

In this model, statistical data can be structured in an easy and transparent manner. Political debates on health services also often follow this structure.

The input of the health-care system is divided into (Schwartz and Busse 2003):

- Patient-side input, that is, the health status of the population as well as its access to care
- Resource oriented input, that is, the input in terms of financial and non-financial resources, such as human resources and infrastructure, as well as organizational structures, responsibilities, and interdependencies between actors and organizations

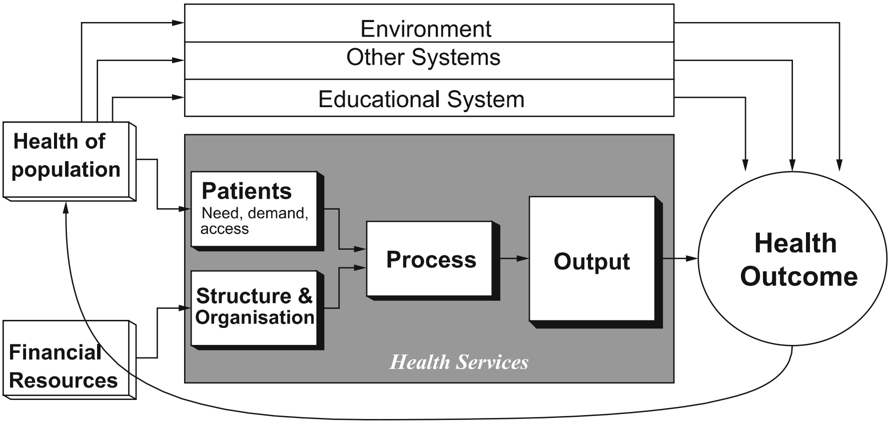


Fig. 24.1 The input-output model (Busse and Wismar 2002)

Throughput forms the center of the model – encompassing all processes of care in a health-care system.

The output of the health-care system is divided into two sequential elements (Schwartz and Busse 2003):

- Direct results of the processes, that is, output measures in the classical sense, also termed intermediary outcomes, for example, the number of cardiac catheterizations performed
- Outcomes in terms of changes in health status, which are often only measurable in the long-term, for example, the mortality avoided by a specific intervention

Common to all models deployed is the problem of causal inference. Although some problems are also encountered by epidemiological research, like the establishment of precedence in time in case-control studies or in historical cohorts (Hill 1965), these problems are much more important in health services research and thus make the latter more prone to biased or confounded results. The main problem is the complexity of the system with multiple interdependencies which result in the dilemma of "before the intervention is after the intervention." A good example is the evaluation of health-care reforms, which often come in a piece-meal fashion and which are only half-way executed before the next reform measures start. Assigning observed changes in an evaluation study to one particular reform package then becomes difficult, and as with all uncontrolled before-and-after studies, the results of such studies have to be interpreted with great caution (Grimshaw et al. 2001). However, in particular, for the evaluation of health reforms, such before-and-after studies are often the only possible research method, as conducting controlled studies is often not feasible.

24.1.3 Level of Analysis

A complementary approach to the one described is the analysis of the level at which processes of care take place (Schwartz and Busse 2003):

- The macro level – consisting of the health system as a whole and national health policy
- The meso level – research focusing on interorganizational structures and processes, for example, between health-care payers and providers, or the relationships between providers in a specific region
- The micro level – analysis of individual care services and technologies

Aday et al. (1998) attempted to match research methods to the level of analysis, illustrated in Table 24.1. In their model, the macro level refers to a population perspective on the determinants of the health of communities as a whole ("health of population" in the model), and the micro level represents a clinical perspective on the factors that contribute to the health of individuals at the system, institution, or patient level (Aday et al. 1998). Their intermediary system level encompasses both the macro and meso level in our model. It refers to the resources (money, people, physical infrastructure, and technology) and the organizational configurations used to transform these resources into health-care services either for the country as a whole (macro level) or within a specific region (meso level).

Table 24.1 Levels of analysis in health services research (Adapted from Aday et al. (1998))

	Level of analysis			
Data sources	Population	System	Institution	Patient
Census	x			
Public health surveillance systems	x			
Vital statistics	x			
Surveys				
Population	x			
Organizations		x	x	
Providers		x	x	
Patients		x	x	x
Insurance records/ administrative data				
Enrollment	x	x	x	x
Encounters		x	x	x
Claims		x	x	x
Medical records		x	x	x
Qualitative studies				
Participant observation	x	x	x	x
Case studies		x	x	x
Focus groups	x		x	x
Ethnographic interviews	x			x

24.2 Methodological Considerations

Generally all types of data can be analyzed for the purposes of HSR. We find experimental data from randomized controlled trials as well as observational data from case-control or cohort studies, registers, or surveys. But many analyses in HSR make use of data from large administrative databases that are abstracted from medical or hospital discharge records, prescriptions, and bills of payments for delivered health services.

24.2.1 Study Designs

24.2.1.1 Randomized Controlled Trials

When alternative approaches to the delivery of health care have to be evaluated, a randomized controlled trial (RCT) is considered the gold standard, that is, the most rigorous method available. RCTs are performed in order to avoid bias and confounding. It is the effect of randomization that provides equality of study and control group in all relevant characteristics except the intervention being tested.

Regardless of this advantage, one has to keep in mind some critical issues when considering the RCT methodology. First, the recruitment of participants (or other experimental units like communities or schools, etc.), who meet all eligibility criteria, may be difficult and expensive. Furthermore, a randomized assignment of treatments to patients may not be feasible by ethical reasons (for instance, if you want to compare a treatment that is widely believed to be efficacious with "no treatment" or with a placebo). The study population is frequently not representative of the target population. Thus, it is true that an RCT has a high level of "internal" validity, as study group and control group are really comparable, but this tends to be connected with a low level of "external" validity which is an important consideration in HSR as it aims to examine effects under actual conditions and not under trial conditions. The results of an RCT refer solely to efficacy (a treatment is called efficacious if the desired effect is obtained under optimal conditions) but not necessarily to effectiveness (a treatment is called effective if the desired effect is obtained under everyday conditions).

Accordingly, the role of RCTs in HSR is more limited than in other areas of health research, and RCTs have only been carried out in certain areas of HSR. For example, the efficacy of cholesterol-lowering treatment in the prevention of coronary heart disease in men with high cholesterol was demonstrated by a multicenter, randomized, double-blind clinical trial, the *Lipid Research Clinics Coronary Primary Prevention Trial* (Lipid Research Clinics Program 1984). As Kelsey et al. (1998) report, the most expensive research study ever sponsored by the US National Institutes of Health – the Women's Health Initiative (Buring and Hennekens 1992) – consists of a series of RCTs to test the hypotheses, whether a low-fat dietary pattern protects against breast cancer and colon cancer, whether hormone replacement therapy reduces risk for coronary heart disease, and whether calcium and vitamin D supplementation protects against hip fractures. Even in the context of evaluating organizational change, RCTs had been carried out. The so-called *Health Insurance Experiment* (Newhouse 1974) was designed as an RCT to evaluate the effect of different levels of cost sharing in health insurance on utilization, expenditures, and health status (for more details, see Sect. 24.3.2). But regardless of such examples, the majority of RCTs are designed to evaluate a new (drug) therapy and are performed in clinical settings (randomized *clinical* trials).

24.2.1.2 Observational Studies

An overwhelming part of HSR is based on observational studies. In a pure epidemiological setting, case-control and cohort studies are used to estimate and evaluate the association between a specific exposure and a specific disease. In addition, exposure and outcome are frequently mapped into binary variables. In HSR, exposures and outcomes have a higher degree of variety than in chronic disease epidemiology.

Typical exposures are

- Conditions that may lead to inequalities in access to care, for example, low-income status, rural area of residence
- Health states that define certain needs for care, for example, mental illness
- Medical interventions, for example, stent implants versus bypass surgery to prevent heart attacks
- Different health-care delivery systems, for example, Health Maintenance Organizations (HMO) versus capitated Preferred Provider Organizations (PPO) versus a traditional indemnity plan with fee for service (FFS) payments
- Programs that aim to improve the quality of care, for example, disease management programs
- Programs to contain the costs of care, for example, drug formularies

Typical study outcomes are

- Access to care, for example, preventive services (vaccination), therapeutics
- Health status, for example, incidence of prespecified (tracer) diagnoses
- Life years gained, that is, reduction of mortality
- Patient reported outcomes, for example, health-related quality of life
- Quality of care scores, for example, measures of the Health Plan Employer Data and Information Set (HEDIS)
- Appropriateness of care measures, for example, an appropriateness evaluation protocol (AEP)
- Cost of care, for example, increases (additional costs or losses) or decreases (savings)

The general limitations of observational studies are dealt with among others in chapters ▸Cohort Studies, ▸Case-Control Studies, ▸Modern Epidemiological Study Designs, ▸Confounding and Interaction and ▸Design and Planning of Epidemiological Studies of this handbook. The absence of randomization gives reason for special concerns, that is, a special type of selection bias, when the goal of the study is to evaluate interventions. Persons who choose a particular intervention – or are advised by a physician to undergo it – are often on a different level of risk for the outcome of interest compared with persons who are not assigned to or did not choose to use this intervention (Selby 1994). Particularly in the case of an open intervention program, persons who pay attention to their health may be more likely to participate and comply with a recommendation (e.g., to undergo a screening examination) than persons who do not (Kelsey et al. 1998).

Case-Control Studies The methodology of case-control studies is treated in great detail in chapter ►Case-Control Studies of this handbook. Because of its obvious merits (a case-control study can be carried out at relatively low cost and comparably quickly), it is used with increasing frequency in HSR, especially in order to asses the adverse effects of drugs and other therapies and to evaluate the efficacy of preventive interventions (Kelsey et al. 1998).

Examples for the application of the case-control approach in HSR are numerous. This includes evaluations of vaccine efficacy and vaccination effectiveness, assessments of medical therapies, of screening programs for cervical, breast, and colon cancers, and of a number of programmatic activities in the community (Armenian 1998).

Cohort Studies Primary data collection for classical epidemiological cohort studies (cf. chapter ►Cohort Studies of this handbook) is relatively rare in HSR compared to chronic disease epidemiology. One reason might be that health systems change fast in small scales, that is, in trends in coding diagnoses, and in large scales as, for example, completely new reimbursement structures like diagnosis-related groups (DRGs) so that information from long-lasting follow-up studies may often be outdated and not worth the expense.

The majority of cohort studies in HSR is based on administrative data collected for purposes other than research and is focused on the outcomes of a medical treatment or a preventive intervention. The outcomes vary and may include mortality, morbidity, functional status, quality of life, costs, and satisfaction with care. The studies frequently use historical cohorts. For example, the investigation of short-term (30-day) and long-term (5-year) mortality in a cohort of members of a large HMO with hip fractures was a historical study that used computer-stored hospital discharge data linked with computer-stored data from death certificates (Petitti and Sidney 1989). A different type of cohort analysis in HSR focuses on the description of changes in symptoms, functional status, or quality of life in patients who undergo a treatment or are the subject of a preventive intervention (Petitti 1998a).

Cross-Sectional Studies If the goal of data analysis is related to health planning or the assessment of needs for services, prevalence rates are often more useful than incidence rates. Cross-sectional studies therefore represent an important tool for health planning and evaluation. In outcome research, the common methodological approach of *variance in practice* (e.g., to assess the quality of medical care or the outcome of the health system of a county) is tightly connected to a cross-sectional design, mostly based on administrative data, with organizations (e.g., hospitals), providers (e.g., surgeons), counties, states, or even countries as the units of analysis.

Cross-sectional studies are also used to establish research priorities based on consideration of the burden of disease (Kelsey et al. 1998). In a study on the prevalence of chronic gynecologic conditions among US women of reproductive age, for example, it was found that the most common conditions were menstrual disorders, adnexal conditions, and uterine fibroids. The results stressed the need for more effective treatments for these disorders and moreover, suggested that more research on their etiology would be highly desirable (Kjerulff et al. 1996).

Cross-sectional studies are of course less useful to examine hypotheses on causal effects mainly because of the lack of knowledge on the temporal sequence of hypothetical causes and potential effects but also because cross-sectional studies include both new and old cases. This results in a case group which has more than its fair share of individuals with disease of long duration because those who die or recover quickly will be underrepresented (Kelsey et al. 1998).

24.2.2 Complex Models for Data Analysis

In several health systems, available claims data are characterized by a longitudinal structure with long strings of repeated measures of health services for individual patients. Such data structures demand analytical designs. To make full use of them, complex longitudinal data analysis techniques must be applied that can handle time-varying exposures, repeated outcomes, and intra-person correlations.

The lack of detailed information on the severity of disease in claims data sometimes is a reason to use case-based study designs, as for instance, case-crossover studies to allow cases to be their own controls (cf. chapter ▶Intervention Trials of this handbook).

Another complicating factor is that observations in health-care delivery systems are often not independent. For many observational studies, the level of observation is a patient (characterized by a vector of patient attributes). A cluster of patients will be seen by the same physician (characterized by a vector of physician attributes) and will therefore experience similar treatment patterns so that their outcomes cannot be expected to be completely independent. Physicians often practice in groups sharing similar practice styles. These groups may practice in a larger health-care delivery system that imposes constraints to treatment choices, for example, drug formularies or payment by capitation (a lump sum per patient), which will make practice styles of groups within a health plan more similar than groups outside the plan. This clustering of observations on multiple levels has led to the adoption of multilevel regression models as standard tools of HSR.

24.2.3 Data Sources

Primary data collection in a randomized controlled trial, a case-control or a cohort study, is certainly an important, although unusual, data source of HSR. Primary data, when used in HSR, are more frequently collected from the general population (or subgroups) by questionnaire. The majority of data that are analyzed in HSR stem from large administrative data bases, as pointed out before.

24.2.3.1 Surveys

Survey research is frequent in HSR. For a detailed description of survey methods, see chapter ▶Epidemiological Field Work in Population-Based Studies of this handbook. On the one hand, it can be used to provide snapshots of the current state of a health-care delivery system. On the other hand, survey subjects often become re-interviewed in regular intervals to form a longitudinal data structure or a panel. Further possibilities to classify surveys are given by the

- Unit of observation (patients, patient-provider contacts or providers)
- Target population (total population or subgroups)
- Type of data collected (interview data, data of medical examination, or both)
- Access to information (personal interview, mail survey, or telephone interview)

A well-known German survey – the *EVaS-Study* (*Study among office-based ambulatory care physicians in the Federal Republic of Germany*, Schwartz and Schach 1989) – was a cross-sectional survey with patient-physician (or patient-office) contacts as units of observation. The concept followed the US *National Ambulatory Medical Care Survey* (*NAMCS*, Tenney et al. 1974) to some extent. The target population was defined by a number of selected regions in Germany, a fixed study period, and the exclusion of a few medical specialties concerning the involved physicians. The data were collected by mail using an induction interview questionnaire, a reporting form, and a final questionnaire. The final data record covers data of the patient as well as data provided by the physician's office (including, e.g., the diagnosis corresponding to the patient's major reason for encounter, the assessment of the severity of the problem, the services delivered, and the duration of the encounter).

The *German National Health Interview and Examination Survey (GHS)*, however, (carried out from October 1997 to March 1999) targeted the general population aged between 18 and 79 years (Bellach et al. 1998).[1] The units of observation had been residents who were interviewed and medically examined. The data are available for research as a public use file. One of the results of this survey, for example, concerned the utilization of medical services available in Germany under statutory sickness fund facilities. About 90% of all Germans had seen their doctor at least once a year. Half of the population had consulted a doctor during the past 4 weeks and, on average, a medical practitioner was consulted 11 times a year (Bergmann and Kamtsiuris 1999). The Robert Koch Institute continued the GHS with the *German Health Interview and Examination Survey for Adults*. The recruitment period for this survey is November 2008 until November 2011 involving a total of 180 cities and municipalities all over Germany (Robert Koch Institute 2008a).

[1] This survey had three predecessors in the years 1984–1986, 1987–1989, and 1990–1991 and was supplemented by the *German National Health Interview and Examination Survey for Children and Adolescents (KiGGS),* carried out from May 2003 to May 2006. The target population were children and adolescents aged between 0 and 17 and living in Germany (Kurth 2007). The Robert Koch Institute continued the KiGGS study by carrying out telephone-based health interviews (Robert Koch Institute 2010).

Another frequently cited German survey is based on a representative, regionally stratified sample of 0.4% of all prescription forms, which are completed by office-based physicians for members of statutory sickness funds. This survey is supported in cooperation by the federal associations of the office-based physicians, the statutory sickness funds, and the free-standing pharmacies. It is carried out each year. The annually published results include an analysis of the sales increase with respect to its components referring to prices, volumes, and structural composition (Schwabe and Paffrath 2010).

The US *Medicare Current Beneficiary Survey (MCBS)* is an example of a survey that is designed as a panel. MCBS began in 1991 as a continuous panel in order to provide a more complete picture of the use of health services, expenditures, and sources of payment for the Medicare population. It is an ongoing computer-assisted personal survey of Medicare beneficiaries residing in the United States and Puerto Rico. Each person is interviewed three times per year over 4 years (regardless of whether he or she resides in the community or a long-term care facility), following a 4-year rotating panel design. The MCBS thus contains four overlapping panels of Medicare beneficiaries. Each year, one panel is dropped from the survey, and a new one is added. This design produces three calendar years of medical utilization data for each sample person. The data are collected over a 4-year period in which sample persons are interviewed 12 times. The first interview collects baseline information on the beneficiary. The next 11 interviews are used to collect three complete years of utilization data. Included are medical expenditure data as well as detailed data on health conditions, health status, use of medical care services, charges and payments, access to care, satisfaction with care, health insurance coverage, income, and employment (Adler 1994). The data are used to produce calendar year public use files on access to care, and cost and use. The nationwide MCBS data are released – as usual for public use files – only under a data use agreement. In addition, requests for regional or supplementing data must include a study protocol with specific justification for the additional data required, along with an identifiable data use agreement (see http://cms.gov/mcbs).

Another excellent population-based US panel, created by the (former) Agency for Health Care Policy and Research and the National Center for Health Statistics, is the *Medical Expenditures Panel Survey (MEPS)* which collects data from several sources to provide a complete picture of the health status and health-care utilization of a random sample of citizens (Cohen 1997).

In addition to other sampling methods, computer-assisted telephone interviews have become more frequently used in HSR. This method has comparatively low costs and guarantees an approximate full coverage of the general resident population in developed countries which have high rates of telephone access. Even unlisted households can be covered by means of random digit dialing.[2] Data are checked for correctness, completeness, and plausibility and stored continuously in the

[2] But the increased use of cellular phones poses a problem, and there is need for research to broaden the approach beyond the restrictions of the conventional telephone network.

course of the interview. Separate steps for data input and examination are not necessary. Germany, for example, started the first *National Telephone Health Survey* in September 2002. About 8,000 German speaking residents aged 18 years and older had been questioned on diseases, health-related behavior, and utilization of the health-care system (Ziese et al. 2003). This survey was supplemented by a regional one in Bavaria (Meyer et al. 2002). The recent telephone survey carried out by the Robert Koch Institute, the *German Health Update*, began in July 2008 and ended in April 2009. Twenty-five thousand people aged 18 and older were selected for an interview. The German Health Update was supplemented by regional surveys in Brandenburg and Saarland (Robert Koch Institute 2008b).

24.2.3.2 Official Statistics

Official mortality and other health or demographic statistics, especially vital statistics (births, marriages, deaths, etc.), have been extensively used in HSR. An early well-known and frequently cited example in the context of equity research is the study on differential mortality in the United States (Kitagawa and Hauser 1973). The comparison of mortality rates and the proportions of money from the national accounts that are spent on health is a popular starting point for health economists in analyzing the efficiency of a particular health system. But official statistics – mostly based on law – resemble routine registries and share their limitations (cf. chapter ►Use of Health Registers of this handbook and Sörensen 2001). The validity of mortality statistics in particular is strongly dependent on the rate of autopsies in a country. In a German survey of institutes of pathology within universities and in community hospitals in the year 2000, a median value of 23.3% and 13.3% of autopsies among hospital deaths was found, respectively. This was considered clearly below the recommended value of 30% (Schwarze and Pawlitschko 2003).

24.2.3.3 Administrative Databases

Administrative data are abstracted from medical or hospital discharge records, prescriptions, and bills of payments. Thus they have several advantages. They are routinely collected data representing the reality of health-care delivery. They need no additional time and money to gain access to large patient populations over long periods of time with repeated recordings of most health-care encounters of each subject. But the advantage of quick and easy access to large and representative populations is counterbalanced by data that may be incomplete and suffer from voluntary and involuntary miscoding. Although the quality of these data appears to improve over time, it has to be kept in mind that the primary reason for creating them was to document medical diagnoses and interventions obtained from medical records and manage the flow of payments for delivered health services obtained from claims data.

Since the advantages, particularly of electronic claims data, are so obvious, researchers try to better understand the consequences of the data limitations and develop analytical methods to adjust for them. Say, for example, the approach of Newhouse and McClellan (1998) used to overcome the typical selection problem, they were confronted with in analyzing the data of catheterization of patients with acute myocardial infarction. As data limitations are unique to each administrative database, a very good understanding of how data were generated is crucial for interpreting analytical findings.

Examples of administrative databases in the USA include the national Medicare and Medicaid databases as well as claims files for privately insured patients or members of a particular health facility. Data from the Medicare program, run by the *Centers for Medicare & Medicaid Services (CMS)*, are confidentiality-protected, longitudinally linked, person-level records that track virtually all elderly US citizens from their 65th birthday onward until death, through geographical moves and changes in providers.[3] The data sets include the types and amounts of health services used (e.g., hospitalizations, office visits, home health care, surgeries, and diagnostic tests), the medical problems being treated (diagnoses), provider characteristics (site of service and physician training), and charges. Information on long-term care services and outpatient prescription drugs, not covered by CMS, is not included (Diehr et al. 1999).

In Germany, administrative databases which do not find their ways into the official statistics or any kind of survey are scattered across the statutory sickness funds or other agencies of social security. Due to comparably strict data protection rules, record linkage is not a common practice in Germany.

From 2004 onward however, the associations of sickness fund physicians have been obliged by law to transfer beneficiary-related billing data, including diagnoses, to the statutory sickness funds.[4] Since then, anonymized/pseudonymized beneficiary-related databases from different research institutes were established, which encompass the data of several insurance companies and are used for purposes of drug safety and health service research (cf. Grobe et al. 2006, 2011; Ihle et al. 2008; Pigeot and Ahrens 2008; Glaeske and Schicktanz 2010; Bitzer et al. 2010; Sauer et al. 2010; Rothgang et al. 2010; Schäfer et al. 2011). In addition, the Federal Insurance Office governs a sample, which is annually drawn from all statutory insured persons, used for the calculations on risk structure equalization.

The abundance of information in claims databases in various states is often overwhelming. Many redundant measures are recorded, and researchers must identify the underlying variables that represent the concepts they want to evaluate. Since researchers on the other side, hate to discard already recorded information, *data reduction techniques* including comorbidity scores or propensity scores are increasingly applied to condense data while preserving information. For a detailed discussion of pharmacoepidemiological databases, we refer to chapter ▸Screening of this handbook.

[3] About 10% of Medicare enrollees are younger, disabled persons, who are tracked from their time of certification.

[4] The hospitals and the pharmacies had started to transfer their beneficiary-related data to the sickness funds long before.

24.2.4 Measurement Error (Misclassification)

In HSR, measurement errors (non-differential and differential as well; cf. chapter ►Measurement Error of this handbook) for all variables of interest are considered to be higher than in a traditional epidemiological setting for several reasons:

- Some data that are collected and stored but are not directly used for reimbursement or other administrative purposes (e.g., job or social status of an enrolled person) are likely to be not up-to-date.
- Diagnostic information on claims is documented to justify reimbursement; a bill for tests to rule out cancer, for example, may contain a diagnostic code for “cancer” even if the tests were negative.
- The information on clinical conditions in administrative data is in the form of diagnoses coded using the International Classification of Diseases (ICD) which is revised from time to time. Currently, ICD-10 (the tenth revision) is in use. Several countries (e.g., the USA) prefer ICD-10-CM, a clinical modification of the ICD. Some diseases such as arthritic or psychiatric disorders are difficult to classify because of lack of clearly defined diagnostic criteria. Less serious or vague conditions have a high probability of inconsistent coding. Regional and temporal variations of coding patterns may additionally reduce the reliability of coded diagnostic information.
- Especially the reliability of ambulatory diagnoses is a major concern. An analysis by the Medicare Payment Advisory Commission – MedPAC (1998) – demonstrated the inaccuracy of outpatient diagnosis coding. For the purposes of the study, MedPAC selected beneficiaries whose Medicare Part B claims in 1994 showed a diagnosis of 1 of 11 serious diseases, then checked for claims for the same diagnosis in 1995. As shown in Table 24.2, the likelihood of a claim in 1995 was only about 50–60% for each of the 11 diagnoses (cf. Newhouse et al. 1997). Part B Medicare covers the costs for medical service by general practitioners,

Table 24.2 Persistence in diagnostic coding of those identified in 1994 (Source: Medicare Payment Advisory Commission 1998, p. 17, Note: Excludes those who died in 1994 or 1995)

Diagnosis on 1994 Part B claim	Percent with Part B claim in 1995
Hypertension	59
Coronary artery disease	53
COPD	62
Congestive heart failure	61
Stroke	51
Dementia	59
Rheumatoid arthritis	55
High-cost diabetes	58
Renal failure	56
Quadriplegia/paraplegia	52
Dialysis	59

for a small selection of pharmaceuticals, for ambulatory treatments in hospitals, and other therapeutic and care services that are not covered by Part A Medicare.

- Moreover, administrative data used to reimburse hospitals or physicians are subject to some problems that are not familiar to epidemiologists, called "upcoding," "coding proliferation," and "gaming." Upcoding of diagnoses to more serious conditions is the process of assigning a diagnosis code or codes to a patient that may maximize the provider's reimbursement (e.g., ischemia to myocardial infarction) as it has occurred with some DRG payment systems (Dunn et al. 1996). Coding proliferation means the increase in the coding of all related conditions affecting treatment. Both types of distortion are relevant sources of measurement errors in HSR. Gaming is a serious problem that is dealt with in Sect. 24.5.2.3 because it cannot be subsumed under the term "misclassification" seamlessly.

In summary, the quality of claims data may be adequate for some purposes, but it is important to remember that claims are generated to justify reimbursement rather than to facilitate research.

24.2.5 Sampling Issues

A considerable part of data analyzed in HSR is based on samples from the population of interest (this is called the "target population" or the "population being sampled"). Sampling can help to save time and money. Sampling may also result in an increase of accuracy of measurement since more effort can be spent on this issue if only a manageable number of units of observation is included. Scientifically sound sampling methods are indispensable tools for designing an efficient sample and to provide consistent and unbiased projections from complex samples. Scientific sampling means *probability sampling*, that is, the probabilities of selection must be under control. Non-random samples based on volunteers or on the judgment of the sampler are not covered by this concept and are not recommended for use in HSR.

The sampling procedure is called *simple random sampling* if each of the possible samples of a given size has an equal chance to be selected. It follows that every one of the sampling units in the population has the same chance of being included in the sample. This is occasionally offered as the definition of simple random sampling (e.g., Kelsey et al. 1998) without realizing that there are other sampling procedures (e.g., systematic sampling, see below), which also have this property (Sukhatme and Sukhatme 1970). Simple random sampling is simple in theory but less so in practice because one needs a complete list of the population to draw the sample.[5] In many instances, it may not be the most efficient method of sampling. Therefore – apart from telephone sampling – it is not much used in practice.

[5] Random digit dialing in order to sample for computer-assisted telephone interviews is considered as a way to handle this problem if the existing lists are not complete and the target population can be accessed by conventional telephone network.

Systematic sampling is a common type of sampling based on selecting every *k*th individual from a list or a file after choosing a random number from 1 to k as starting point. It is based on a fixed rule and is not limited to selection from an actual file. Thus, selection of all those born on the (randomly chosen) third day of any month or of everyone whose social security number ends in (the randomly chosen digits) 17, 48, or 76 is similar to systematic sampling procedures yielding approximately 3% samples.[6] Because of these properties, systematic sampling is often simpler to administer under field conditions than simple random sampling. But systematic sampling has a severe handicap: differently from other sample designs, it is impossible to estimate the variance from one single sample. For an unbiased estimate, you need repeated sampling. Several (biased) approximations are used in practice to estimate the variance. One of these consists in treating the systematic sample as if it was a random sample of *n* units (Sukhatme and Sukhatme 1970).

When the population can be divided into *strata* in such a way that each stratum is more homogenous than the population as a whole, one can reduce the sampling error compared to a simple random error. Examples of variables that are often used for stratification are region, age, sex, race, and socioeconomic status. Following a *stratified sampling* design, a separate sample is drawn from each stratum, and the results are then appropriately combined in the analysis.

Cluster sampling has contrasting properties compared to stratified sampling. It is a simple random sample applied to groups of population members (*clusters*) that usually leads to a substantial loss of precision. But because of operational improvements of access to the units to be selected, one often achieves a heavy decrease of collecting costs and thereby an increase in precision per unit of cost. Examples of clusters that could be sampled are hospital wards, villages, schools, families, etc. If clusters are positively correlated within themselves, that is, they have a high positive *intraclass-correlation coefficient (ICC)*, indicating more homogeneity than would result from chance alone, cluster-sampling variance will be larger than simple random sampling variance. This is a situation frequently observed in real life. As ICCs are positive in most cases, simple sampling variance can grossly underestimate the true cluster-sampling variance. The ratio of the latter to the first-mentioned variance is called the *cluster effect*.

In a *multistage sampling design*, stratification and clustering may be combined on several stages of the sampling procedure forming a complex random sample. Stratified sampling, for example, may be used to ensure that schools are represented in the sample according to different socioeconomic areas in a large city, and cluster sampling of classrooms within the selected schools might then be employed for efficiency.

The analysis of data from a complex sample procedure that includes cluster sampling requires a sound knowledge of sampling theory or statistical advice. There

[6] Of course, the choice of the sampling scheme has to be relevant to the population being sampled. A population that is not completely covered by social security would be unsuitable for sampling by means of social security number.

are a series of textbooks on sampling theory (e.g., Hansen et al. 1953; Kish 1965; Stuart 1968; Sukhatme and Sukhatme 1970; Cochran 1968; Levy and Lemesshow 1991), and many handbooks or textbooks on statistics contain at least a chapter on sampling theory (e.g., Kendall and Stuart 1958; Kahn and Sempos 1989; Krishnaiah and Rao 1994; Voß 2003). The use of a special software (e.g., Sudaan) or a special module of one of the common large statistical packages (e.g., SAS, Stata or SPSS) is inevitable. They allow for variance weighting in the statistical procedures to adjust for the specific sampling design. Otherwise, as cluster-sampling variance may be many times larger than the variance calculated by assuming a simple random sample (Abraham 1986), and the analysis can result in severely misleading conclusions about the significance of the study findings.

24.2.6 Confounding and Risk Adjustment

For general principles of control for confounding, see chapter ►Confounding and Interaction of this handbook. Health services researchers tend to summarize methods to adjust for confounding under the term "risk adjustment." With respect to the large databases analyzed, standardization and multivariate modeling are more frequently used to control for confounding than the traditional approach of stratification.

Any level of comparison can be affected by confounding. This includes the mapping of health-care needs, the evaluation of clinical strategies and programs, studies of the effectiveness of quality improvement initiatives, or the evaluation of cost containment measures. Typical confounders are age, sex, ethnicity, income, smoking, or other risk variables. In outcome studies, confounding is a major concern because of differences in severity of illness and comorbidity.

The most frequent approach to control for confounding in HSR is to include the potential set of confounders as predictors in the regression model (cf. chapter ►Regression Methods for Epidemiological Analysis of this handbook) to predict the outcome of interest:

- Ordinary least squares (OLS) regression when the outcome has a continuous distribution (ideally a normal distribution) as, for example, the logarithm of costs
- Poisson (or negative binomial) regression when the outcome is described by counts (as, e.g., the number of hospital admissions in a specified year)
- Binomial (logistic) regression when the outcome variable is binary, indicating, for example, the occurrence of disease or death

When a large database is used for the analysis, comorbidity is often taken into account by including a lot of so-called dummy, that is, binary 0/1 variables in the regression model that indicate the presence (or absence, respectively) of each out of a list of classified comorbidities. When using samples of small or moderate size, this approach may not be possible. In this case, a premodeled aggregated index of comorbidity can be included in the analysis (Schneeweiss et al. 2001).

Including a potential confounder variable in the analysis requires its storage in the database. This is crucial whenever "secondhand" data are analyzed. Especially in a country like Germany where record linkage is not a common practice, one cannot expect to have full control over all relevant confounders, and many socioeconomic characteristics are available only in survey research.

Omission of one or several confounders usually leads to a violation of the assumptions underlying the estimation procedure in the OLS regression model

$$Y_i = \beta_1 x_{i1} + \beta_1 x_{i2} + \ldots\ldots.. + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots., n,$$

as the k predictors x_{ij}, $j = 1, \ldots, k$, and the random errors ε_{i}, $i = 1, \ldots, n$, are no longer uncorrelated where n denotes the number of subjects, Y_i the response variables, and β_j the regression coefficients.

In this situation, the introduction of one or more of so-called *instrumental variables* can help to establish a consistent estimation of the interesting effects (that is, β_j). This has been a well-known technique of econometric analysis (IV-technique) for over half a century which is described in almost every econometric textbook (e.g., Greene 2003). But it is very rarely applied to HSR problems because a sound econometrical or statistical background is needed. Instrumental variables should be correlated with the predictor variable as much as possible, and at the same time, they should be (at least asymptotically) uncorrelated with the random errors. The skill of the technique consists in finding such variables which are already in the database or could be added to it at a tolerable cost.

An illuminating example of the usefulness of IV-technique is presented by Newhouse and McClellan (1998) who explain the instrumental variable convincingly as a device that achieves a pseudo-randomization. The authors analyzed the effect of catheterization and associated revascularization of acute myocardial infarction on mortality in the years following treatment. For IV-estimation of this effect, they used the differential distances of the patients' place of residence to the (nearest) catheterization, revascularization, and high-volume hospital as instrumental variables.

24.3 Demand, Need, Utilization, and Access to Health Care

A main focus of HSR is the assessment of demand, need, utilization, and access to health services, which represent closely related but distinct fields of investigation. In the input-output model, they represent the endogenous, risk-related input, which is, among others, determined by population health – representing the exogenous risk-related input in the model.

Demand is a general economic concept, which can be defined as the "quantity of a good buyers wish to buy at a conceivable price" (Begg et al. 1997a). Demand for health and health care is in many respects different from demand for other goods and services. The demand for health care is a derived demand as health care is not sought in itself but as a means to improve one's health or to prevent its deterioration. Health care itself is indeed often rather unpleasant (McPake et al. 2002). Health is not something that can be traded, and both health and health care are surrounded by uncertainty. What people want in essence from health care is to buy access to care in case they need it, that is, insurance (McPake et al. 2002). Another aspect is that health is both a consumption good and a capital good (McPake et al. 2002). Especially politicians and health-care funders often focus on the consumption side and neglect the potential of investing in health as a durable good which is an important prerequisite for economic growth.

The notions of need, utilization, access, and the relationships between them will be discussed in more detail in the following sections.

24.3.1 Assessing Health Needs

The concept of "need" for health and health care links directly to population health. Initially it appears simple and is often used by politicians in health policy discussions but quickly becomes complicated and is therefore avoided by many analysts (Kindig 1997). Instead many policy analysts, in particular, in the United States, prefer to use the economic demand and supply framework, where it is assumed that if someone "needs" something, he or she will express this desire by purchasing the item that is needed in the marketplaces, and as a consequence, supply will increase (Kindig 1997).

An alternative concept of need as the "capacity to benefit" has been proposed by Williams (1974) and Culyer (1993). Their concept also goes beyond the perspective of an initial baseline level of health because unhealthy individuals and populations cannot be said to need more health care without regard to their potential for improving their health status (Kindig 1997). Capacity to benefit also rules out health services which might be desired by individuals or providers but which do not make a positive contribution to health-adjusted life expectancy (Kindig 1997). It also goes beyond a mere epidemiological description of health needs in terms of ill-health or shortcomings in care in a specified population as it incorporates the notion of effectiveness of the intervention. Some authors use the terms "felt need," "normative need," and "expressed need." The "felt need" reported by patients is often substantially different from the "normative need" as judged by health professionals. "Expressed need" represents the need expressed by action, for example, visiting a doctor (Wright 2001).

Three main approaches to health needs assessment exist (Wright 2001):

- Epidemiologically based needs assessment – combining epidemiological approaches, such as specific health status assessments, with assessment of the effectiveness and possibly cost-effectiveness of interventions
- Comparative needs assessment – comparing levels of service receipt between different populations
- Corporate needs assessment – canvassing the demands and wishes of professionals, patients, politicians, and other stakeholders

In practice, comprehensive health needs assessments often combine all three approaches. Practical applications are manifold, such as to highlight areas of unmet need and to provide a clear set of objectives to meet these needs, to decide how to use resources to improve the local population's health, and to influence policy, interagency collaboration, or research and development priority setting (Wright 2001).

A good example for a health needs assessment is a study contrasting the epidemiological need for carotid endarterectomy with actual service provision in an English region (Ferris et al. 1998). The authors estimated the need for a carotid endarterectomy on the basis of demographic and epidemiological data, assuming that the rate of endarterectomies in their region should match the rate of patients with symptomatic carotid disease – the patient group for whom carotid endarterectomy is proven to prevent strokes. Based on estimates of the incidence of transient ischemic attacks ($77/10^6$ population/year) and minor strokes ($76/10^6$/year), they calculated that the need for endarterectomy was $153/10^6$/year, which contrasted with operation rates of $35/10^6$/year in 1991–1992 and $89/10^6$/year in 1995–1996. The ratio of use to need was 0.47 (95% confidence interval (CI) from 0.4 to 0.54), which was far from being satisfactory. Furthermore, they noted a disconcerting variation in the use to need ratios between districts, ranging from 0.28 (95% CI: 0.19–0.38) to 0.81 (95% CI: 0.62–1.06), and lower use to need ratios in elderly and female patients – indicative of inequity in access in relation to need. The epidemiological needs assessment was supplemented with a corporate needs assessment comprising interviews with vascular surgeons and a joint purchaser-provider workshop. These indicated that the low operation rates were primarily due to low rates of referral for diagnostic assessment by general practitioners. The variation between districts partly reflected the concentration of services – districts with a high use to need ratio tended to have one of the main provider sites. This study clearly demonstrates the usefulness of such research in identifying the main levers for improvement of current service provision – in this case, raising awareness for the clinical indications for carotid endarterectomy in general practitioners, in particular, those located in rural areas without access to a local vascular surgical service.

24.3.2 Assessing Utilization and Access to Services

Studies assessing the utilization of services attempt to improve our understanding of who uses health-care services and why (Black 1997). In the previous sections, it has already become clear that many factors determine whether a patient utilizes a health-care service, among them whether the patient suffers from a condition for which an effective intervention is available and whether he or she demands that service. Three other common determinants of utilization have become apparent in the example of the health needs assessment for carotid endarterectomy in England – clinician's judgment, distance from facilities, and sex. If a general practitioner does not consider that referral to a specialist is necessary, it is unlikely that the patient will end up having the procedure nonetheless – resulting in unmet need and underutilization of health services. Other important factors influencing utilization and access are patients' knowledge and the cost of services to the patient.

A well-recognized example for the influence of sex on utilization is higher rates of appendectomy in women than in men (Black 1997). After the primary assertion that appendicitis is more frequent in women was not supported by evidence, the more likely explanations were as follows: Appendicitis-like symptoms are more common in women, probably arising from ovarian dysfunction; in some cultures, young women prove their independence by undergoing an operation; operation rates are dependent on the availability of services – the more surgical services are available, the higher the sex difference in operation rates (Black 1997).

The influence of clinicians' judgment on health service delivery has been investigated in a series of studies comparing hospital care and related costs between Boston and New Haven in the United States – two cities with similar demographics where most hospital care is provided by university hospitals (Wennberg et al. 1987, 1989). However, in 1982, expenditures per head for inpatient care in Boston were about twice as high as in New Haven ($889 vs. $451) (Wennberg et al. 1987). The excess utilization in Boston compared to New Haven totaled $300 million and 739 hospital beds per year. In a subsequent study in 1985, Wennberg and colleagues showed that the variation in operation rates between the two areas did not result in statistically significant differences in mortality between the two cities (Wennberg et al. 1989). The authors concluded that the lower rate of hospital use in New Haven was not associated with a higher overall mortality rate in the populations concerned and consequently that hospital care was overutilized in Boston and not underutilized in New Haven.

The influence of geographical and financial barriers to access is also well documented. Black and colleagues (Black et al. 1995) attempted to identify the reasons for geographical variation in the use of coronary revascularization in the United Kingdom in a cross-sectional study. They found considerable variation in revascularization rates between districts, which arose from differences in supply factors, notably the distance to a regional revascularization center and the existence of a local cardiologist. The level of coronary heart disease mortality in the population and the lack of use of alternative treatments not only failed to explain the observed variation but was inversely associated with the rate of revascularization (Black et al. 1995). This inverse relationship between need and provision of care has been observed in many settings and has been termed the "inverse care law" by Julian Tudor-Hart (Tudor-Hart 1971, 2000). It has to be kept in mind that in measuring the utilization of health-care facilities, only those patients are counted, who have surmounted barriers to access – be it long distances, fear of an operation, lack of public transport, waiting lists – and are thus biased (Schwartz and Busse 2003). These barriers to access also exist in countries which grant a legal right to health care to every citizen, in particular, among socially disadvantaged groups in society. These differences in access to care are even more pronounced in countries without such a right to health care and where direct financial barriers to care exist on top of other barriers to access.

The effect of financial barriers to accessing health services has been studied in many countries at different levels of development. The most famous study is the RAND Health Insurance Study (see also Sect. 24.2.1.1) on the effect of cost-sharing measures on utilization (Newhouse 1974; Newhouse et al. 1981; Brook et al. 1983). Between 1974 and 1977, about 2,000 non-elderly families were randomly assigned to different insurance plans. Participants were assigned to either prepaid group practices or to one of 14 fee-for-service insurance plans, which varied in their coinsurance rates and in maximum spending per family and year (Newhouse 1993). The authors found that the more families had to pay out of pocket, the fewer health-care services they used. Families on the plan with the highest coinsurance (95%) up to a $1,000 limit on annual family expenditure reduced expenditure by 25–30% compared to a plan which was free to the family (Newhouse 1993). Interestingly, the use of all types of services, whether physicians, hospitals, pharmaceuticals, dental, or mental health services, fell with cost-sharing to a similar degree, except hospital admissions of children which did not respond to plan (Newhouse 1993). While the reduced utilization had no negative effect on the health for the average person, health among the "sick poor" – the most disadvantaged 6% of the population – was adversely affected (Newhouse 1993). Especially the poor who began the experiment with elevated blood pressure had their blood pressures lowered more on the free care plan than on the cost-sharing plans, and mortality rates predicted on the presence of major risk factors were 10% lower among those insured on the free plan (Newhouse 1993). Free care at the point of delivery also improved both near and far corrected vision, increased the likelihood that a decayed tooth would be filled, and the prevalence of anemia among poor children was lower (Newhouse 1993). All observed adverse health effects of cost-sharing hit the poor and less educated disproportionately more.

A number of other factors can limit access to services, in particular, gender, age, professional status, race, and religion (Schwartz and Busse 2003). The discrepancy between need in terms of ill-health and capacity to benefit from intervention and utilization is a commonly used measure of equity.

24.4 Financial Resources, Structure, and Organization of Health Services

Health-care financing can be described from different perspectives:

- The first one looks for the ultimate sources of funding. Here, intermediary sources of financing (government, social security funds, private social insurance, and private households above all) have to be tracked back to their origin.
- The second one, commonly used in National Health Accounts, aims at a breakdown of expenditure on health into the complex network of third-party payments plus the direct payments by households or direct funders (OECD 2000).
- The third one focuses on the allocation of the available resources. Health planning and management of health care among others includes the continuous task of distributing the financial resources, for example, to distinct segments in the natural history of disease (as reflected in prevention, cure, rehabilitation, and care), to alternative treatments for a specific disease, to different regions, to various groups of providers, or simultaneously with respect to some of these categories.

When comparing the developed countries with respect to financing from the first perspective, we can find two marked types of health services systems: systems that are funded by taxes and typically have a National Health Service (NHS) – so-called NHS-type or Beveridge-type systems – and systems that are predominantly financed by contributions of employees and employers (so-called Bismarck-type systems). In the real world, we find, of course, mixed systems including all the common forms of public and private financing. The UK is generally considered to be the classic example of an NHS-type health service system, and the German health system, as established by Bismarck in the late nineteenth century, naturally acts as the model of all Bismarck-type systems.

Health service systems also vary considerably in the overall health insurance coverage rates. In all EU member states, nearly 100% of the population has coverage of some type of public or private health insurance, whereas in the USA, 46.3 million persons (corresponding to 15.4% of the population) had no coverage at all during the entire year 2008 (U.S. Census Bureau 2009), which constitutes an important barrier to access as seen above.

International comparisons of health-care spending are hampered by a variety of definitions and classifications, used by the national statistical agencies, resembling the different organizational structures of health-care delivery (van Mosseveld 2003). To improve comparability of health, accounting data Eurostat, the statistical office of the European Union (EU), and the Organization for Economic Co-operation and Development (OECD) jointly developed a conceptual basis of rules for the statistical reporting of health accounts together with EU member states (OECD 2000). This so-called *System of Health Accounts (SHA)* corresponds to the new German health expenditure data system which was developed simultaneously (Brückner 1997; Statistisches Bundesamt 2000a, b). The SHA includes expenditures of private households, is consistent with the System of National Accounts (SNA) methodology (OECD 1993), and covers three dimensions: functions of health care, providers of health care, and sources of funding. Core of the SHA is a newly developed *International Classification for Health Accounts (ICHA)* that is based on a three-digit code.

The ICHA reflects the potential variations among health systems in structure and organization of health care and the share of work between the various providers. One remarkable distinction between health services systems, when looking at the organization, refers to outpatient care. In many countries, patient-physician contacts take place only in offices of general practice. Contacts with medical specialists are limited to hospital visits no matter whether the patient has been admitted to the hospital or not. But there are other countries, among them Germany, where outpatient specialized curative care is predominantly provided by office-based specialists.

There are many additional differences in the structure and organization of health services systems in spite of the fact that in all developed countries, patients with a specific health condition receive more or less the same treatment, provided that quality standards are observed. The package of activities in health care seems to be stable over the health systems, while the providers are different. International studies of health-care systems based on comparable data sets are rare. One of the outstanding examples is the WHO/International Collaborative Study of Medical Care Utilization – WHO/ICS-MCU (Kohn and White 1976). The data collection process of this carefully designed and methodically ambitious study included a cross-national multilingual household survey (by personal interviews) and standardized forms relating to health services resources and organizational factors. The study included almost 48,000 respondents representing over 15 million persons in 12 study areas scattered over Europe and America. One of the striking results of the WHO/ICS-MCU study was that study areas with the highest estimates of societal interest in health were also the areas with the lowest totals for per capita health expenditure and for health expenditure as a percentage of national income.

24.4.1 Allocation of Resources

Health-related decision makers in government, regional authorities, insurance companies, or other institutions are faced with the task of allocating resources. Examples are allocating research funds to different areas of HSR, Medicaid funds to treatments, Medicare funds to HMOs, or global funds to local authorities. Regional allocation is a main concern of all NHS-type health-care systems. Risk-adjusted capitation payments to insurers or to providers is a related topic that has been discussed extensively in several non-NHS-type countries (e.g., the USA, the Netherlands, and Germany).

There is no unique method for resource allocation analysis. Different levels and variable purposes of resource allocation analysis require different methods:

- For economic evaluation, for example, of drugs, surgical procedures, other types of clinical interventions, or of community intervention programs, limits on health-care resources mandates resource-allocation decisions guided by considerations of cost in relation to expected benefits (Weinstein and Stason 1977).
- The UK approach of weighted capitation has become the principal method of allocating health-care finance to regions (Rice and Smith 1999).
- Risk-adjusted capitation, whereby capitated payments are adjusted to reflect the expected cost of individual enrollees, is commonly based on multivariate regression models to predict health-care expenditure (Van de Ven and Ellis 2000).

While economic evaluation are mostly based on RCTs and observational studies (mainly on effectiveness and costs), risk-adjusted capitated payments and formulas of weighted capitation are generally based on official statistics (e.g., census or mortality statistics) or large samples from administrative databases. For an example, see Sect. 24.4.1.2.

24.4.1.1 Economic Evaluation, Especially Cost-Effectiveness Analysis

Economic evaluation can be defined as the comparative analysis of alternative courses of action in terms of both their costs and consequences (Drummond et al. 1997). There are four main types of economic evaluation (see Table 24.3).

Table 24.3 Types of health economic evaluations (Source: Epstein and Sherwood (1996))

Type of analysis	Assumption/question addressed
Cost-minimization	The effectiveness (or outcome) of two or more interventions is the same. Which intervention is the least costly?
Cost-effectiveness	The effectiveness of two or more interventions differs. What is the comparative cost per unit of outcome for the intervention?
Cost-utility	The question is the same as for cost-effectiveness analysis. The outcome is a preference measure that reflects the value patients or society places on the outcome.
Cost-benefit	The effectiveness (or outcome) of two or more interventions differs. What is the economic trade-off between interventions when all of the costs and benefits of the intervention and its outcome are measured in monetary terms?

But in practice, most of the health economic evaluations apply the cost-effectiveness methodology. *Cost-benefit analysis* is rarely used in public health and health-care settings because of methodological difficulties to measure the value of human life and low acceptability of its results on the side of health policy decision-makers and health professionals. *Cost-utility analysis (CUA)* is considered by many to be a subtype of *cost-effectiveness analysis (CEA)* where the effectiveness measure includes societal or individual preferences for the outcomes – a customary effectiveness measure in CUA is the quality-adjusted life year (QALY) (cf. Sect. 24.6.1.2 and chapter ▶Descriptive Studies of this handbook) – compared to natural units as effectiveness measure in CEA, for example, life years gained or mmHg blood pressure reduction. Most of the important methods and concepts applicable to cost-effectiveness studies are also applicable to cost-utility and cost-minimization studies. Cost-of-illness studies may be identified as a fifth type of economic study in HSR. Their goal is to estimate the total societal costs of caring for persons with a specific illness compared to persons without this illness, irrespective of any intervention. Such studies are carried out to demonstrate the (relative) burden of illness. They are not full economic evaluations because alternatives are not compared (Drummond et al. 1997).

One limitation that is common to all types of economic evaluation arises from the difficulty in obtaining a true estimate of costs, particularly in a health-care or public health setting where high proportions of fixed costs and little flexibility in changing the labor pool are typically found (Petitti 1998b).

A common understanding of cost-effectiveness claims that one of the three criteria has to be met (Doubilet et al. 1986). First, an intervention is cost-effective when it is less costly and at least as effective as its alternative. Second, an intervention is cost-effective when it is more effective and more costly, but the added benefit is "worth" the added cost. Third, an intervention is cost-effective when it is less effective and less costly, and the added benefit of the alternative is not "worth" the added cost.

Cost-effectiveness is measured as a ratio of cost to effectiveness. Two concepts to calculate this ratio should be distinguished (Detsky and Naglie 1990): An average cost-effectiveness ratio is estimated by dividing the cost of the intervention by a measure of effectiveness. An incremental cost-effectiveness ratio is an estimate of the cost per unit of effectiveness of switching from one intervention to another. In estimating an incremental cost-effectiveness ratio, both the numerator and denominator of the ratio represent differences between alternative interventions (Weinstein and Stason 1977). Often the terms "marginal" and "incremental" are used interchangeably in the literature, although *marginal costs* are strictly speaking the costs of producing one extra unit of output, whereas *incremental costs* usually refer to the difference, in cost or effect, between the two or more programs being compared in the economic evaluation (Drummond et al. 1997).

Estimating *average cost-effectiveness ratios* can be useful for service planning and for resource allocation decisions between very different health programs, for example, an influenza vaccination program and liver transplantations. However, for resource allocation decisions between interventions for the same disease, for example, two different antihypertensive drugs, *incremental cost-effectiveness ratios* should be used. The importance of using incremental cost-effectiveness ratios for decision making in some settings is best illustrated with the example of the sixth Guaiac stool test to screen for colorectal cancer, which had been endorsed by the American Cancer Society and which has later been shown to have an incremental cost of $47 million per case detected compared to an average cost of $2,451 per case detected (Neuhauser and Lewicki 1975).

The unspecified implicit alternative to an intervention is usually doing nothing. But doing nothing has costs and effects that should be taken into account in the analysis (Detsky and Naglie 1990). Furthermore, explicit declaration of "doing nothing" as the alternative intervention helps to frame discussions of the desirability of the intervention (Petitti 1998b).

Costs seem to be a straightforward notion, well understood by everybody. But actually, it is a rather complex term that consists of various components: direct costs, indirect costs, and intangible costs. Costs that are directly related to an intervention (and to side effects and other consequences) are summed up to the total of *direct costs*. By *indirect costs,* health economists understand the monetary value of lost wages and productivity due to morbidity and death of a person affected. *Intangible costs* refer to consequences that are difficult to measure and value, such as the value of improved health or the pain and suffering associated with illness or treatment (Drummond et al. 1997). The rationale of economic evaluation is based on the concept of *opportunity cost*, that is, the benefits forgone by not deploying resources for the next best alternative use.

As costs are seen differently from different *perspectives* (e.g., perspectives of health insurers, corporations, hospitals, physicians, and patients), it is also important to define a cost perspective in CEA and state it explicitly (Petitti 1998b). A common goal in CEA is the societal perspective so that the total costs of the intervention to all payers for all persons are included in the analysis.

Costs and benefits, after all, must be discounted before comparing them by calculating the ratio of cost to effectiveness. *Discounting* is the usual procedure in economics used to determine the present value of future money. This analysis gives a greater weight to costs and benefits the earlier they occur. High positive discount rates favor alternatives with costs that occur later or benefits that occur earlier. This clearly favors curative versus preventive health programs. In the business world, there is no fixed rate of return on investment, and the use of a private sector return rate for public sector program cost may not be correct (Sudgen and Williams 1990). Most published CEA in developed countries use discount rates between 3% and 5%. An expert panel commissioned by the US Public Health Service, based on the "shadow price of capital," recommended using a discount rate of 3% for economic evaluation in the public health sector (Gold et al. 1996). Whether benefits should also be discounted, and if so at what rate, is highly controversial.

Estimates of benefits and costs in a CEA may be uncertain because of imprecision in both underlying data and modeling assumptions. Therefore major assumptions should be varied and the net present value and other outcomes computed repeatedly to determine how sensitive outcomes are to changes in the assumptions. This so-called *sensitivity analysis* is typically the last step in a CEA. A sensitivity analysis varying the discount rate from 0% up to 7% should always be done (Gold et al. 1996).

As illustrated by Oregon's Medicaid reform efforts in 1990/1991, CEA or other types of economic evaluation cannot be used as sole basis for allocating scarce resources because the question of equity and ethical issues are not addressed by this method. In Oregon, CEA was only 1 of the 13 factors used to prioritize funding of services for the poor (Petitti 1998b).

Since Sir William Petty found out in 1667 that public health expenditures to combat the plague would achieve a benefit-cost ratio of 84 to 1 (Fein 1971), numerous studies of economic evaluation have been carried out, most of them using ratios of cost to effectiveness. The list of interventions that were economically evaluated within the last 10 year spans from influenza vaccination of healthy school-aged children (White et al. 1999) to colonoscopy in screening for colorectal cancer (Sonnenberg and Delco 2002), and preoperative autologous blood donation (Etchason et al. 1995) to reducing the population's intake of salt (Selmer et al. 2000).

24.4.1.2 Weighted Capitation in NHS-Type Health Systems

The central aim of weighted capitation is to distribute a global health budget between geographical areas in accordance with population needs and thus provide equal opportunity of access for equal needs. Currently used formulas of weighted capitation can be described as a modified age standardization of health-care expenditure. The UK Resource Allocation Working Party (RAWP) originally recommended in 1976 that the resources for the hospital and community health services (HCHS) be distributed on the basis of population size, weighted by age and sex, the need for health care, and the costs of providing services (Carr-Hill 1989; Advisory Committee on Resource Allocation (ACRA 1999)). Standardized mortality ratios (SMRs) were used as a proxy measure for relative needs. However, this had been criticized for failing to fully reflect the demand for health-care resources related to chronic disease and deprivation.

In 1995, a new weighted capitation formula for HCHS was introduced. This comprises an age index (based on estimates of national resources spent per capita in eight age groups) and an "additional need" index (additional to that accounted for by demographic variables). The need weighting index takes the form of four indices for acute, psychiatric, non-psychiatric community, and community psychiatric services, which are based on 1991 small-area census socioeconomic variables. It is derived from an empirical model that identified its need indicators as those census-derived health status and socioeconomic variables which, having been adjusted for the independent effects of supply, were most closely correlated with the national average pattern of hospitalization (Carr-Hill et al. 1994).

For all its merits, however, this formula, also called English formula, and the models on which it is empirically based have been criticized. The fundamental criticism relates to the use of utilization-based models to assess need for health care, which implies that historical patterns of service uptake between different care groups (as revealed by utilization) are appropriate (Mays 1995).

Against this background, some scientists pleaded for a radically new approach to health resource allocation, one that distributes NHS resources on the basis of direct measures of morbidity rather than indirect proxies such as health service utilization or deprivation. The Welsh steering group on allocation (Townsend 2001), for example, recommended the use of a morbidity-based budgeting approach. In a study of target allocations for the inpatient treatment of coronary heart disease in a sample of 34 primary care trusts in different areas in England, it was shown that a morbidity-based model would result in a significant shift in hospital resources away from deprived areas, towards areas with older demographic profiles and toward rural areas (Asthana et al. 2004). In the discussion of their findings, the authors concluded by calling for greater clarity between the goals of health-care equity and health equity.

Up from the year 1999, the Advisory Committee on Resource Allocation (ACRA) of the Department of Health took care of the further development of the Weighted Capitation Formula. ACRA picked up the above mentioned critique and incitations and learnt to distinguish between the two objectives "equal access to healthcare" and "reduction of health inequalities." Moreover, an index to estimate unavoidable labor costs, known as the staff Market Forces Factor (MFF), was included in the formula.

The actual formula for 2009–2010 and 2010–2011 allocations is built up by three components: hospital and community health service, prescribing and primary medical services. The corresponding weights are 76%, 12%, and 11%. Each component reflects both the additional needs and the unavoidable costs (the prescribing component does not have an adjustment for unavoidable costs since the prices of drugs do not vary by geographical location.)

The components again are set up by two indices: one (model-based computed) aiming at utilization of services and the other one aiming at health inequalities uses disability-free life expectancy as its measure, combining 2005 life expectancy data with 2001 limiting long-term illness data, and so capturing morbidity as well as mortality (Department of Health 2008).

24.4.1.3 Risk-Adjusted Capitation

When competition is an essential component of a health-care system, it is a widespread belief that capitated payments create incentives to contain costs and to compete on quality. But they also create undesirable incentives for risk selection ("cream skimming"), that is, to attract profitable patients (or enrollees) and to avoid unprofitable ones, and to decrease service intensity.

Risk adjustment is an important tool to reduce cream skimming while encouraging desirable cost and quality competition. This method controls for confounding (comorbidity above all) by calculating the expected health-care costs (or some other measure of an outcome) for members of health plans or insurance companies. This control is realized either by stratification (*cell-based approach*) or by multivariate modeling (*regression approach*). In many developed countries around the world, health-care organizations have established some sort of risk adjustment procedure for resource allocation. Examples exist for the following countries: Austria, Brazil, Canada, Chile, Germany, Hong Kong, Israel, the Netherlands, Spain, Switzerland, Taiwan, and the USA. Most of the risk adjustment procedures are based on a regression model to predict future health expenditure. Models differ with respect to the set of included predictors, the procedure of grouping diagnostic information, and the populations used for calibration.

The exclusive reliance on risk-adjusted capitated payments has been criticized, for example by Newhouse et al. (1997), who pointed out that the common risk adjusters (predictors of cost) are not likely to reduce risk selection problems to negligible levels. This concern was confirmed by a study of Shen and Ellis (2002) who examined the maximum potential profit that plans could hypothetically gain by using their own private information to select low-cost enrollees when payments are made using one of four commonly used risk adjustment models. Their findings – based on simulations using a privately injured sample – suggested that risk selection profits remain substantial (Shen and Ellis 2002).

Against this background, it was recommended to move the financing of health services to partial capitation payments. Partial capitation for an individual enrollee combines capitation methods and some reflection of that person's actual use of services, for example, a fee for service payment. Partial capitation would reduce plans' incentives to select good risks – the intent of risk adjustment – and also reduce the financial incentive to underserve or stint on care (Newhouse et al. 1997).

The General Form of Regression-Based Risk Adjustment Model Frequently used are regression models with untransformed costs as the dependent variable, estimated by ordinary least square (OLS). The standard assumptions of that type of statistical model (namely, a normal distribution, homoscedasticity, and independent observations) are not satisfied sufficiently by utilization data, but for predicting future costs, the model has shown to work about as well as more complex models in real situations (Diehr et al. 1999).

Occasionally a two-part model is applied: One equation predicts the probability that a person has any use, and a second equation predicts (on a log scale) the level of use for users only. In a two-part model, the regression coefficients of the first equation are estimated by logistic regression analysis and those of the second equation by OLS regression. Two-part models tend to meet the assumptions better than one-part models and provide insight into the utilization process, but they are not recommended when the goal is to predict future costs because transformations cause complications in this context (Diehr et al. 1999).

The list of possible predictors of a model for risk-adjusted capitation includes age, sex, and other demographic or socioeconomic variables, as well as binary variables to indicate that a person has been assigned to a diagnosis belonging to a special group from a system of diagnostic groups or has received a drug prescription belonging to a special group from a system of drug categories. To incorporate information on morbidity, some models use hospital diagnoses alone, while others use both inpatient and outpatient diagnoses. It has, however, to be noted that previous utilization is a strong predictor of a future one. This means that the costs in 1 year heavily depend on utilization in the year before for chronically ill patients. Thus, as in any regression analysis, it is important not to control for this variable lying on the causal pathway (Diehr et al. 1999).

The estimated regression coefficients ("regression weights") refer to the so-called calibration population. For diagnosis-based models, generally this is also the population used to establish the diagnostic classification system, the "grouper." Recalibration of a model without a refinement of the grouper therefore may lead to biased estimation. Generally the models are calibrated prospectively (that is, the data of the predictor set refers to the previous year, while the cost data refer to the actual year), but in order to evaluate the predictive power, current calibration (both types of data refer to the same year) has been performed as well.

The standard summary measure of model performance in prediction is R^2, the percentage of the total variance of the dependent variable that is explained by the model. Usually the values of R^2 in prospectively calibrated models do not exceed 20%. Newhouse et al. (1989) used theoretical and empirical arguments to estimate that the maximum possible R^2 in the context of utilization data is about 15% for total expenditure (prospectively modeling).

In addition to the grouper and the regression module, any risk adjustment methodology finally requires a module that links the estimated costs to the payment system or the resource allocation procedure, respectively, that is, a mechanism that controls the way payments or the allocation of resources is based on the predicted health-care expenditure.

Risk Adjustment in the US Setting Up to 1999, Medicare paid the HMOs 95% of the *adjusted average per capita cost (AAPCC)*, an estimate of the expected cost of treating Medicare beneficiaries in the fee-for-service sector in each local area. The AAPCC methodology adjusted for differences between the HMOs enrollees and fee-for-service users with respect to age, sex, welfare status, and whether or not they were in a nursing home (Ellis et al. 1996).

Since its implementation in 1985, the AAPCC had prompted concern about its fairness and accuracy, and it was shown that only about 1% of total variance of the cost of treatment was explained by this concept (Newhouse 1986; Ash et al. 1989). Against this background, the *Health Care Financing Agency (HCFA)* sponsored the development of alternative approaches that include diagnostic information as predictors in the regression-based risk adjustment model, among them the *Diagnostic Cost Groups (DCG)* family and the *Adjusted Clinical Group (ACG)* methodology (Ingber 1998). In the years 2000–2003, AAPCC has been stepwise replaced by the *Principal Inpatient Diagnostic Cost Group* model *(PIP-DCG)* which uses sociodemographic variables and hospital diagnoses to predict next years cost (Pope et al. 2000). In view of the widespread concern about the quality of ambulatory diagnoses, the DCG family was supplemented in 2001 by a model that uses outpatient pharmacy data, grouped into 127 mutually exclusive categories, instead of ambulatory diagnoses (Zhao et al. 2001). From 2004 onward, the CMS/HCC-model,[7] a 100% comprehensive risk adjustment scheme (using full encounter diagnostic data) has been implemented to adjust Medicare capitation payments to private health-care plans for the health expenditure risk of their enrollees (Pope et al. 2004).

Medicaid supported the development of the *Chronic Illness and Disability Payment System (CPDS)* which groups the Medicaid beneficiaries according to a hierarchical diagnostic classification system (Kronick et al. 1996). CPDS, which later on was reconstructed and recalibrated to predict expenditures also for Medicare beneficiaries, has now been established in several US states (Kronick et al. 2002).

Risk Adjustment in a Bismarck-Type European Setting In European countries with a predominating Bismarck-type organization of health services, we find competition among all insurance companies and even among the statutory sickness funds. The main goal of risk adjustment (better: risk equalization) in these settings therefore is to reduce risk selection by the sickness funds and to establish a fair system of income-related contributions. HSR has played a major role in designing and reforming these systems.

Like in the USA, the starting point of risk adjustment in Europe has been set by models based on age, sex, and other sociodemographic variables. The Netherlands, for example, started in 1992 with a prospectively used age- and sex-based model. In 1995, region and disability were included as predictors, and a “high-risk pool” was established in addition. Since 2002, dummy variables were added to the model that indicate prescriptions of drugs falling into 1 out of 13 mutually exclusive categories, the *Pharmacy-based Cost Groups (PCGs)*, each of them closely related to a serious chronic disease (Lamers 1999). From 2004 onward, the Dutch risk-equalization methodology had been further supplemented by an inpatient DCG module that uses hospital diagnoses only.

[7] CMS: Centers for Medicare & Medicaid Services; HCC: Hierarchical Condition Categories

In 1994, Germany introduced a retrospective risk-equalization procedure among statutory sickness funds which was based on the following variables: age, sex, and two dummy variables indicating invalidity or disability pension and the entitlement for sickness allowance. The procedure was also designed to adjust for different incomes because the beneficiaries pay income-related contributions. The largest share of the risk-adjusted financial transfers between sickness funds (up to 60%) results from differences in per capita income of the beneficiaries. From 2002 onward, the German risk-equalization methodology has been extended. First, a retrospective "high-cost pool"[8] was established, and second, a dummy variable was added to the set of risk adjusters indicating that a beneficiary is registered in an accredited disease management program. From 2009 onward, a risk-equalization procedure based on a DCG/HCC module has been established, using inpatient and outpatient diagnoses with respect to 80 (by an expert panel) selected diseases.

24.4.2 Evaluating Effects of Organizational Characteristics and Change

As health services systems in the developed countries tend to go through one reform after another and are more or less continuously exposed to change, evaluation is a permanent task of HSR. But the preconditions do not favor the establishment of scientifically sound designs of research. Experimental designs are extremely rare. The above-cited RAND Health Insurance Study on the effect of cost-sharing measures on utilization is one of the most famous exceptions. In some circumstances, it is even difficult to implement a quasi-experimental design including a control group. Particularly in countries like Germany, where benefits and programs are uniform but the organizational responsibilities are widely scattered over local authorities and institutions, evaluation research is very complex.

Suggested by the structure of available data, perhaps the most frequently used quasi-experimental design for analyzing aggregated annual data in the context of program evaluation is the time series experiment. It can be characterized by a periodic measurement process on some group and the introduction of an experimental change X into this time series of measurements O_i. Adapting a diagram by Campell and Stanley (1966), the time series design can be outlined as follows (whereby the number of observations before or after X, occurring here in year five, may be smaller or larger as in a real problems):

$$O_1, O_2, O_3, O_4, X, O_6, O_7, O_8, O_9$$

The main problem of (internal) validity inherent in a time series design is revealed by seeking likely alternative explanations of the shift in the time series

[8] The high-cost pool consists of insured with high cost in the past year (above a fixed threshold) which are shared by all statutory sickness funds.

other than the effect of X. This problem, of course, could be settled to a great extent by establishing a suitable control group (comparison series) that shares all intervening factors except X with the study group.

A natural approach for analyzing data from a time series design is *segmented* or *piecemeal regression* (e.g., Neter and Wasserman 1974). This method is appropriate when the considered response variable has a linear trend over the range before X (segment one) followed by another linear trend over the range after X (segment two). The year which divides the segments (year five in the above diagram) is known as the join point (or break point). When the hypothetical change of trend line refers only to the slope and not to the intercepts (that means no discontinuity between the both lines), the regression equation for analyzing data from a design as diagrammed above can be specified as follows:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1 x_2,$$

where Y is the response variable, x_1 is the year ($x_1 = 1, 2, 3, 4, 6, 7, 8, 9$) and x_2 is a dummy variable indicating that the year is greater than 5. The parameter β_2 measures the difference in slopes between the lines. If the same trend continues from the first segment to the second segment, then $\beta_2 = 0$. The β_j are estimated, and the hypothesis $\beta_2 = 0$ is tested, using standard procedures in regression. It is easy to expand segmented regression to more than two segments and even to allow for discontinuities between segmented regression lines. Autocorrelated errors or heteroskedasticity can be handled by using standard techniques (e.g., Greene 2003).

A good example for applying this type of analysis to evaluate the impact of a program on the basis of aggregated data is the study of the effect of a regionalized perinatal care program in North Carolina (established in 1965) on perinatal and postneonatal mortality (Gillings et al. 1981). A similar – but due to autocorrelation, more complex – analysis was carried out to evaluate the effects of patient-level payment restrictions for prescription drugs under Medicaid in the years 1981–1983. By this analysis, supplemented by survival analysis to measure the rate of admissions to hospital and nursing homes, it could be shown that the decline in the use of drugs after the cap (a limit of three paid prescriptions per month) had been associated with an increase in rates of admission to nursing homes (Soumerai et al. 1987, 1991).

Regardless of these examples, there is only limited use of OLS regression models for evaluation because of several restrictions. First, costs or the logarithm of costs or other continuously distributed responses are only one type of outcome measure used for evaluation. Counts of specific events, for example, contacts, prescriptions, hospital admissions, etc., and binary response variables like death, accident, or first occurrence of a specific disease are equally important – in some situations, even more important measures. Second, if individual longitudinal data are available, to make full use of the data structure, the model used should be able to handle clustered (correlated) response data, arising from repeated measurements and time-varying covariates. *Generalized linear models* (Nelder and Wedderburn 1972;

McCullagh and Nelder 1983) and the related *generalized estimating equations* (GEEs) (Liang and Zeger 1986; see also chapter ▶Generalized Estimating Equations of this handbook) form the methodological framework of an advanced approach of statistical modeling for evaluation. Consistent parameter estimates in these models are achieved by maximizing likelihood or quasi-likelihood functions using some sort of Gauss-Newton algorithm. Several of the common packages of statistical software, among them SAS, provide corresponding procedures.

The standard model to analyze count data, for example, is the *Poisson regression model*, which is a non-linear regression model that can be formulated as a generalized linear model. Poisson regression is robust insofar as consistent estimation of the regression coefficients does not require that the dependent variable is Poisson distributed. Only a correct specification of the conditional mean is required (Cameron and Trivedi 1998). But Poisson regression is prone to overdispersion. Therefore the condition that the variance equals the mean has to be relaxed by introducing a dispersion parameter that must be estimated as well. Otherwise, testing hypotheses on the regression coefficients could yield misleading rejections of null hypotheses (Cameron and Trivedi 1998).

Poisson regression can also be used to analyze correlated counts from repeated measurements. The within patient correlation is then estimated in the framework of GEEs, whereas the effects of the covariates can be modeled as a generalized linear model. For example, the introduction of reference pricing for angiotensin-converting enzyme (ACE) inhibitors for patients of 65 years of age or older in British Columbia, Canada, in January 1997, was evaluated by using such a type of analysis (Schneeweiss et al. 2002). Several covariates were included in the model, among them age, sex, the adjusted household income, and a chronic disease score computed from prescription medications for every quarter and treated as a time-varying covariate. This ambitious study was based on computerized administrative health databases covering a large proportion of the population, including all types of claims, hospital admissions, admissions for long-term care, diagnoses, and the medications, dose, and dispensed quantity of all prescriptions. Even the deaths within the study cohort were included.

A similar analysis has never been done in Germany, though reference-based prices (RBP) for the beneficiaries of the statutory sickness funds were established in 1992/1993. For reasons of privacy and data protection, cross-institutional linkage of existing scattered administrative databases on drug utilization, ambulatory diagnoses and medical services, and hospital data on an individual level need extensive data protection procedures in Germany. Thus the effects of RBP has to be evaluated on the basis of aggregated data. But any conclusions on the overall economic and public health impact, if obtained solely on the basis of aggregated data, are distorted because of the introduction of fixed drug budgets and the effects of the reunification of Germany (among other confounders) that both took place in the beginning of the 1990s, more or less simultaneously with RBP (Schneeweiss et al. 1998).

Sometimes one has to balance the advantage of using individual longitudinal data – without having a control group – against the advantage of having a control group at the price of rather limited capacities of analysis based on aggregated data. For example, in a study of the effect of premium rebate to reward low utilization of services for beneficiaries of one statutory sickness fund in Germany, the effect on expenditure mainly was analyzed using a long time series of aggregated data together with a control series. In a second step, this analysis was combined with an examination of the effects on non-monetary measures of utilization based on short time series of beneficiary-related data that were primary collected by the sickness fund in order to support administration of premium rebate (Schäfer and Nolde-Gallasch 1999).

24.5 Process of Health Care: Effectiveness, Appropriateness, and Quality

Research on the process of health care considers questions like "Which services are provided in which quantity, by whom, where, and how (Schwartz and Busse 2003)?" The production of health care is a complex result of financing arrangements and of demand- and supply-side factors. The interaction of these different factors is not well understood. An interest in investigating these questions arouse after substantial and unexplained variation in procedures, and hospital admissions were observed between similar hospitals. Some examples for these variations have already been presented in Sect. 24.3.2, for example, the Boston-New Haven Study (Wennberg et al. 1987, 1989). These studies demonstrated the importance of supply-side factors on utilization patterns and frequency, if patient-related factors are controlled for, such as age, sex, case mix, and socioeconomic status. The supply side of a region is primarily described by the density of physicians, hospital beds, and the availability of medical technologies. However, most studies do not analyze the effects of these provider or supply-side characteristics on the health status of the populations concerned (Brook and Lohr 1985). More refined supply-side characteristics determining the use of services comprise provider payment mechanisms, experience and sex of health professionals, organization and equipment of physicians' practices, size and type of hospital, as well as referral patterns between different providers (Schwartz and Busse 2003). "Self-referral" of patients has been identified as an important determinant of small area variation in the use of medical technologies (Childs and Hunter 1972). Self-referral describes the phenomenon of providing expensive medical technology, for example, X-ray examinations, for patients in general practitioners', physicians', or orthopedic surgeons' practices without referring the patient to a radiologist. In comparisons between countries with a comparable standard of health care, the possibility of self-referral for X-ray examinations compared to countries with X-ray examinations exclusively provided by a radiologist increases the overall rate of X-rays by a factor of 4 (Busse 1995). Within a country, differences in examination frequency between doctors with the possibility of self-referral compared to doctors who have to refer patients to a radiologist yield comparable results. In Germany, the rates for X-ray examination for patients with chronic pain were increased by a factor of 2.7, the rates for abdominal ultrasound for patients presenting with gastrointestinal symptoms by a factor of 3.0 in practices with a possibility of self-referral compared to practices who had to refer their patients to other practices (Busse 1995). Of course this observation is linked to the method of physician remuneration. It is a phenomenon which is primarily observed in countries with fee-for-service remuneration such as the United States and Germany. The effects of the structure of financial incentive systems and resulting overutilization on a system level tend to be underestimated. In Germany, fee-for-service remuneration combined with the possibility of self-referral and the widespread practice of non-radiologists to provide X-ray examinations in their practices resulted in 1,655 X-ray examinations being performed per 1,000 inhabitants in 1997 (Deutsche Röntgengesellschaft 2002) – about twice the rate observed in other European countries. It is estimated that unnecessary X-ray examinations during the last decades now cause around 2,000 incident cases of cancer in the country annually (Berrington de Gonzales and Darby 1994).

The extreme variation in health service provision raises the question whether diagnostic and therapeutic procedures are appropriately used in the process of care. To judge whether a procedure is appropriate, knowledge about the effectiveness of the procedure for certain indications or clinical presentations is required. However, this is not the case for the majority of indication-procedure pairs.

24.5.1 Assessing Effectiveness and Appropriateness of Care

In general, the effectiveness of a health-care professional or service is the degree to which the desired outcomes are achieved (Gray 1997). However, the proposition that an intervention is effective implies that there is only one outcome of care and only one objective in the design of that intervention – which is rarely the case (Gray 1997). In addition to a number of beneficial outcomes of care, such as lower mortality and morbidity, the possibility of harmful effects of care has to be considered. Effectiveness research attempts to answer questions, such as "What is the right thing to do?" or "What care confers significant health benefit for a given clinical situation?" (Scott and Campbell 2002).

Another frequently used concept is that of efficacy, which is the impact of an intervention in the best possible circumstances (Gray 1997). These can be achieved in randomized clinical trials (RCTs). However, the reverse conclusion that RCTs always produce efficacy results is not true, as not all RCTs satisfy high quality standards. The distinction between efficacy and effectiveness is important, as the latter represents the impact of an intervention under routine care conditions. The difference between the two concepts in terms of health status outcomes is illustrated in Fig. 24.2 using the example of complications of radical prostatectomy. Data on effectiveness in the example are derived from a meta-analysis of Medicare routine data; efficacy data are taken from a meta-analysis of RCTs.

An important aspect of both efficacy and effectiveness is that they apply to groups of patients. However, the impact of an intervention on the health status of an individual depends to a large extent on individual factors. To answer the

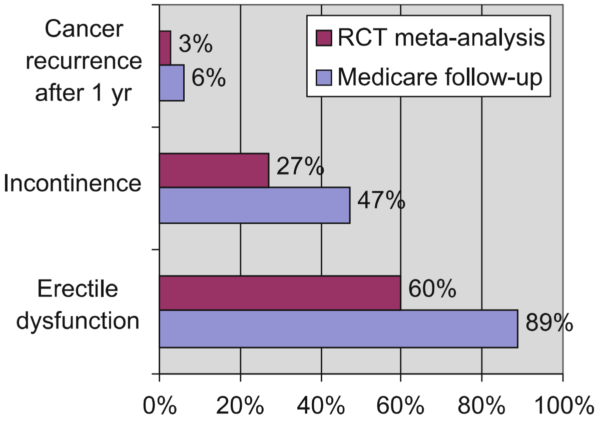


Fig. 24.2 Comparison of effectiveness and efficacy using the example of radical prostatectomy (Adapted from Fowler et al. (1993))

question of whether the most appropriate care was provided given the clinical circumstances is the realm of appropriateness research. An intervention can be considered appropriate, if the expected health benefit exceeds the expected negative consequences by a large enough margin to justify performing the procedure rather than other alternatives (Herrin et al. 1997).

Appropriateness research also addresses the questions of overuse, underuse, or misuse of interventions (Scott and Campbell 2002). We have already discussed the effect of provider remuneration systems and organizational features of a health system on utilization rates. Another major determinant of utilization is clinical judgment. A historical study on clinicians' judgment variation is a study from New York in the 1920s, in which one thousand 11-year-olds had their throats examined (American Child Health Association 1934), cited by Black (1997). In 61% of these children, a tonsillectomy was performed. Of the remaining 39%, the examining doctor thought half would require a tonsillectomy. The half with healthy tonsils were examined by another doctor, who thought that half of them required surgery. The healthy children were again examined by yet another doctor, who declared that half of them required a tonsillectomy, which means that after four examinations only 65 out of 1,000 children would have escaped with their tonsils intact (Black 1997).

During the last 10 years, a number of studies using the RAND/UCLA appropriateness method have been performed. This basically consists in collecting an expert opinion on the appropriateness of an intervention using the Delphi technique. The experts are asked to judge a number of possible indications (that is, case descriptions with clinical information including comorbidity, age, and sex) on whether a specific intervention was appropriate, inappropriate, or even harmful in these cases. In a second step, these judgments are applied to real patient groups in order to determine in how many cases the procedure was appropriate or inappropriate. Most studies using the RAND/UCLA appropriateness method reported appropriate interventions in the range of 50–85% of all interventions performed. Non-US expert groups consistently thought that more procedures were inappropriate compared to US expert groups. This has to be kept in mind when interpreting results from appropriateness studies. An example is a review of the appropriateness of coronary angiography, upper gastrointestinal endoscopy and carotid endarterectomy performed on 4,564 patients in the USA in 1988. The review which was based on a literature review and on an expert panel consensus concluded that, respectively, only 77%, 76%, and 36% of these procedures had been appropriate (Brook et al. 1990). Inappropriate care is more than a nuisance. It can be harmful to health. For example, in a prospective observational study on congestive heart failure admissions, 7% of admissions were found to be the result of improper medical treatment, including fluid overload, procedures, and misuse of drugs. Hospital mortality for this group of patients was 32% compared to 9% in patients without inappropriate treatment (Rich et al. 1996).

24.5.2 Assessing Quality of Care: Clinical Practice Performance

Measures of clinical practice performance continue to be under constant discussion, particularly when they are published routinely, as is the case, for example, for hospitals in the United States and the United Kingdom. There is broad agreement on the dominant paradigm, established by Donabedian (1980), of measuring quality of clinical care in terms of *structure*, *process*, and *outcome*, but each category has advantages and disadvantages which must be assessed in relation to the type and the speciality of the service (e.g., inpatient vs. outpatient care or surgery vs. drug therapy), the condition being treated (e.g., diabetes, hypertension, acute myocardial infarction, or mental disorders), the case mix of the patients, the role of the comparative information, and a variety of other context variables (Shojania et al. 2001).

The type of assessment of clinical practice performance heavily depends on the perspective on quality. Blumenthal (1996) distinguishes four main perspectives on quality: the health-care professional perspective, the perspective of health-care plans and organizations, the purchaser perspective, and the patient perspective. Health-care professionals tend to emphasize technical excellence and the characteristics of interaction between patient and professional (Donabedian 1988). Health-care plans and organizations place greater emphasis on the general health of the enrolled population and on the function of the organization (Leape 1994). Purchasers, of course, additionally incorporate the price and the effectiveness of the delivery of care. Taking into account the preferences and values of patients leads to a definition of quality that emphasizes outcomes such as functional status, morbidity, mortality, or quality of life and encompasses satisfaction with care (Petitti and Amster 1998).

24.5.2.1 Indicators of Structural Quality

Structural measures characterize the resources in the health system. They describe the setting in which care occurs and the capacity of that setting to produce quality (Donabedian 1980; Brook et al. 1996). Quality assurance programs and organizations such as the Joint Commission on the Accreditation of Health

Care Organizations (JCAHO) and the National Committee on Quality Assurance (NCQA) in the USA or the associations of statutory health insurance physicians in Germany rely on structural measures (as listed below) to infer quality and confer accreditation on this basis.

For providers, structural measures include professional characteristics like speciality or board certifications, etc. For hospitals, they include ownership, number of beds, teaching status, licensure status, availability of sophisticated technologies, qualification of personnel, and other organizational factors for inpatient care (e.g., staff-to-patient ratio, closed intensive care units, dedicated stroke units, or the presence of a clinical information system). One frequently used structural measure of quality is patient volume (Shojania et al. 2001). The growing use of this indicator reflects an extensive literature, which documents superior outcome for hospitals and physicians with higher patient volumes for certain indications and procedures (e.g., Luft et al. 1979; Hannan et al. 1989; Phibbs et al. 1996; Thiemann et al. 1999).

When using structural indicators to measure quality of care, the implicit assumption is that structure affects outcome. This is certainly true for the compliance with minimum standards of structure (e.g., rules for hygiene in operating rooms). But on higher levels of structural quality, the link between structure and outcome is less clear (Shojania et al. 2001). For example, specialist care as a quality measure not always results in better outcomes. This is demonstrated by the findings that even cardiologists fail to provide proven therapies to many eligible patients with acute myocardial infarction (Brand et al. 1995). These findings promote the case to measure the processes of health-care delivery directly instead.

24.5.2.2 Indicators of Process Quality

Process indicators permit a glimpse into the inside of the care-delivering units, allowing measurement of the care patients actually receive. They measure the net effect of physicians' clinical decision making. Clinical choices about the use of surgery, medication or diagnostic tests, admission to a hospital, and length of stay account for a large proportion of the costs of services and of outcomes experienced by the patients. Sometimes generic process measures are used (e.g., number of prescriptions, average length of stay, or day case surgery rate). But mostly they are specific to specialities and certain conditions (e.g., antibiotics within 8 h for patients with community-acquired pneumonia, prophylaxis for venous thromboembolism, or beta-blockers for patients with acute myocardial infarction).

Process measures can be reported for individual physicians, groups of practitioners, for hospitals, hospital units, or hospital trusts, or for the entire system of care. They are favored by providers to indicate quality because they are directly related to what providers do. Frequently they are derived from evidence-based clinical guidelines and facilitate individual physician quality improvement. If proven diagnostic and therapeutic strategies are monitored, quality problems can be detected long before demonstrable outcome differences occur (Brook et al. 1996).

Even so there are some arguments against process-based measurement of the quality of care. First, process measures are not necessarily good predictors of outcome, and allocating resources to processes which do not affect outcomes may increase cost without producing any improvement in health (Ellwood 1988). Moreover, collecting process data may be a comparatively elaborate procedure. Finally, it may not be possible to achieve consensus on the recommended process for many clinical problems (Petitti and Amster 1998).

24.5.2.3 Indicators of Outcome Quality and Adjustment for Case Mix

The quality-relevant health outcomes have been described as the "five Ds" – death, disease, disability, discomfort, and dissatisfaction (Elinson 1987), or, more positively turned, when measuring quality of care health outcomes could be summarized as survival, states of physiologic, physical, and emotional health, and patient satisfaction (Lohr et al. 1988). Broader definitions of outcomes include psychosocial functioning, quality of life, resource utilization, and costs of care (Iezzoni 1994).

The use of outcome measures to assess the quality of clinical performance has been criticized for several reasons. First, even for common conditions, it may take years to detect differences in outcomes between groups of patients (Palmer 1997). Moreover, such differences may not be under the control of providers but reflect, among others, patient factors, variations in admission practices, or chance rather than differences in quality of care (Shojania et al. 2001). Many outcomes (e.g., mortality) are rare, and comparisons of quality based on such outcomes often have low statistical power (Brook et al. 1996).

Considerable concern is related to perverse incentives for "upcoding" and "gaming" (McGlynn 1998), whereby gaming means a change of treatment to more expensive forms which are frequently more stressful for the patient and result in a reduced quality of care (e.g., a surgical procedure instead of a drug prescription, an inappropriate hospitalization, or a short hospital admission for a marginal diagnosis).

Incentives for gaming may arise from the criteria used to define target patient populations. For example, restricting inpatient mortality to deaths that literally occur in the hospital allows hospitals to lower their mortality rates simply by discharging patients to die at home or in other institutions (Jencks et al. 1988). Additionally, the incentive for physicians or hospitals to avoid caring for sicker patients remains a substantial concern for outcome-based performance measurement (Hofer et al. 1999). Proliferation of diagnoses related to comorbidity or coding of diagnoses related to severity of illness (upcoding) was observed after the introduction of the prospective payment system for HMO-enrolled beneficiaries of Medicare (Keeler et al. 1990).

The most important concern of research related to outcome-based quality measures focused on the development of case-mix adjustment models for hospital mortality rates. Case-mix adjustment and risk adjustment are based on similar methods, but they use different data sources: Case-mix indices are based on medical records from hospitals or physicians, while risk adjustment is based on administrative data, for example, from health insurances. Models that have originally been designed to predict financial rather than clinical outcomes (see Sect. 24.4.1.3) did not perform sufficiently well in this context because hospital data differ significantly in structure from administrative data. Progress has been made in adopting models to identify the case mix of a group of patients by focusing on specific subgroups of patients instead of overall hospital mortality and using clinical rather than administrative data (Iezzoni 1994).

24.5.3 Examples for Performance Assessment

24.5.3.1 Comparison of HMOs Based on a Performance Indicator System

The federal Centers for Medicare and Medicaid Services (CMS), formerly known as the Health Care Financing Administration (HCA), and the private sector in the USA have supported the development of several performance indicator systems in order to compare the quality of care delivered by HMOs. Perhaps the most popular system, the Health Employer Data and Information Set (HEDIS), was introduced in 1993 and was revised in 1995 and again in 1997. It can be considered as the model for many other performance measurement efforts (Petitti and Amster 1998).

HEDIS was designed by the National Committee for Quality Assurance (NCQA) to evaluate several aspects of health plan performance including clinical quality of care, access to care, satisfaction with care, utilization of services, and the financial performance of the HMO. The clinical quality-of-care indicators included in HEDIS were chosen to address aspects of the care process for which there was strong evidence in the literature to support the relationship between medical care process and desired outcomes (Petitti and Amster 1998). These included indicators for low-birth-weight babies, childhood immunization status, breast cancer screening, eye exams for people with diabetes, and beta-blocker treatment after heart attack.

24.5.3.2 Hospital Ranking

In NHS-type countries, the assessment of the quality of clinical performance focuses on the health of the general population and the function of the health-care system with a special concern on inpatient care. For example, in the United Kingdom, in order to rank hospitals, the publication of clinical indicators in the form of so-called league tables has a long history. As far back as 1983, a set of performance indicators was published covering five areas, one of which was clinical activity. Since then, the set of indicators has been revised several times (British Medical Association 2000). Currently, the published UK league tables are compiled by the Dr Foster organization (separately for England, Wales, Scotland, and North Ireland). They are based on the Department of Health's Hospital Episode Statistics (HES) data and data collected through questionnaires. The indicators fall into five broad categories: standardized mortality rates, waiting times and volumes, staff-to-bed ratios, and – for England only – patient and staff satisfaction and other rating-based scales (clean hospital, good food, etc.). The mortality rates are standardized for age, sex, length of stay, and type of admission (elective or emergency admission). SMRs are calculated for each of 80 ICD9 three-digit primary diagnoses (accounting for 80% of all in-hospital deaths) cited in the final episode of care (Dr Foster 2004a, b). The league tables are criticized for several reasons. The first concern refers to an insufficient control for case mix relating to severity, comorbidity, deprivation, and the availability of places for people to be discharged to nursing homes or hospices. The second refers to the use of HES data, which are based on finished consultant episodes (the NHS's measure of hospital activity), whereas no conversion to hospital spells is provided (HESs are not designed to collect detailed clinical data). Third, the primary diagnosis has been questioned, as diagnostic criteria change. Finally, the focus on inpatient mortality is considered as a shortfall because an increasing proportion of deaths occur outside the hospital (Jacobson et al. 2003). Dr Foster continued to publish the league tables regardless of this critique, and in the Editors' letter of the 2009 report, you can find the following statement: "Over the years, the report has remained a constant as an independent, authoritative guide to hospital care written for the patient, the politician, the civil servant, the manager and the clinician." (Bedford and Kafetz 2009).

In the United States, an annual index of hospital quality ("America's Best Hospitals") is published by *U.S. News & World Report*. This hospital ranking methodology was devised in 1993 by the statistics and methodology department of the National Organization for Research (NORC) at the University of Chicago. The ranking is based on reputation, mortality, and other factors. The reputational scores of a hospital are based on a survey. The index is designed to be used by patients who are looking for the best hospital to treat their health problems. Since 2005, the annual rankings of "America's Best Hospitals" are produced by the Social, Statistical, and Environmental Sciences Division of RTI International (North Carolina).[9] RTI produces hospital rankings with components representing three key aspects of care: structure, process, and outcome. These components are combined to give an overall score for each hospital in twelve medical specialties. For four additional specialties, scores were based on original survey data alone. The mortality score (outcome) is adjusted for case severity. The severity adjustments were derived using the All Patient Refined Diagnosis Related Group (APR-DRG) method designed by 3M Health Information Systems. The APR-DRG adjusts expected deaths for severity of illness by means of principle diagnosis and categories of secondary diagnoses (RTI 2009).

Beginning in 2007, U.S. News & World Report also began publishing separate annual rankings of "America's Best Children's Hospitals." Like Best Hospitals, the Best Children's Hospitals rankings reflect the interrelationship among structure, process, and outcomes. Most structure and outcome data were obtained directly from children's hospitals using the Pediatric Hospital Survey data submission form which is hosted by RTI. In 2010, children's hospitals were evaluated in ten pediatric medical specialties (RTI 2010).

In Germany, there is no published ranking of hospitals except for some studies of limited impact. A methodology of hospital ranking based on routine data of sickness funds and patient questionnaires with respect to total hip replacement was published

[9] This trade name has ist roots in the Research Triangle Institute (RTI), which was established by the universities located in the Triangle's three cities Raleigh, Durham, and Chapel Hill in North Carolina.

(Schäfer et al. 2007; Bitzer et al. 2007) but until now not fully established for public use. Nevertheless, there are some websites which deliver information on quality of hospital performance (structure, process, and outcome). This information is based on data collected by questionnaires from the insured of statutory sickness funds, on data from the quality reports of the hospitals, established by law in 2004, and – for selected diagnoses and procedures (tracer) – on routine inpatient care data of one large statutory sickness fund. For the latter (risk adjusted) approach (cf. Heller and Günster 2008), the hospitals are rated in the following categories: “above average,” “average,” and “below average.”

24.5.3.3 Physician Profiling

The Physician Payment Review Commission of the American Medical Association (AMA) defines physician profiling as “an analytical tool that uses epidemiological methods to compare practice patterns of providers on the dimensions of cost, service use, or quality (process and outcome) of care.” Profiles can be developed for an individual physician, a group of physicians, or physicians within a hospital or managed care plan. They can be broken down by geographical area, speciality, type of practice, or other characteristics. Profiling can focus on many different types of outcome or resource measures. Those resources may be defined globally (e.g., overall charges/costs for the care of a person or group of persons) or they may represent certain subcategories of services (e.g., laboratory, x-ray, physician services, or pharmaceuticals). Profiling is usually applied to compare resources used by cohorts of patients to get a sense of whether their providers do or do not practice efficiently. Even when profiles are not used to modify payment, they may be used to select or reject providers or to determine appropriate patient caseloads for salaried practitioners (Tucker et al. 2002).

The core element of any profiling methodology is risk adjustment by calculating an SMR-like figure, that is, a ratio of observed to expected values of the considered measure. The expected value is adjusted with respect to age, sex of the patients, and to the diagnostic groups that were assigned by the used grouper. Most of the sellers of common models of diagnosis-based risk adjustment (e.g., ACG and DCG) offer the use of their predictive models for profiling.

Profiling may serve as a tool to feed information on care back to the physicians. Also, managed care organizations as a whole have had considerable experience with profiling in order to monitor plan activity. For example, profiling reports, adjusted for case-mix, can be used to distribute bonus or set aside funds that are marked to recognize how well resource management goals are achieved among managed care providers (Tucker et al. 2002). A review of profiling in practice is given by Sutton (2001).

In 1996, the Commonwealth of Massachusetts became one of the first states to implement a comprehensive physician profiling program available to consumers over the Internet. Many other states have adopted similar systems since. In the beginning of the year 2001, physician profiles were available in 30 states, with legislation pending in eight others (Sutton 2001).

In Germany, profiling of physicians has been based on crude measures. Up to now, they have been compared with the average of the regional group of physicians belonging to the same speciality, but a medium-term change to the risk-adjusted profiling system, mandated by law, is scheduled.

24.6 Outcomes of Health Care

24.6.1 Assessing Output and Outcomes of Care

A frequently cited definition of outcomes was given by Donabedian (1985): "Outcomes are those changes, either favorable or adverse, in the actual or potential health status of persons, groups or communities that can be attributed to prior or concurrent care."

The most conventional method of measuring the health status of populations is by means of vital statistics, including statistics of birth and death. Disease-specific incidence rates, cause-specific mortality rates, or other population-based indicators are extensively used to assess the health status of communities, counties, or health systems in general (see Sect. 24.6.3). For example, the Centers for Disease Control (CDC) established a set of 18 population-based health status indicators in 1991 for use at all administrative levels in the United States (Freedman et al. 1991).

Vital statistics may be considered as the key feature of outcome research to study health care and the effect of intervention on a broad range of outcomes, both humanistic and clinical (Petitti 1998a). As population-based measures of health and methods of adjustment are dealt with in chapters ▶Rates, Risks, Measures of Association and Impact and ▶Confounding and Interaction of this handbook and in several sections of this chapter, in the following, we focus on further approaches to measure health status used in outcome research, including patient-based outcomes measurement, adjusted life expectancy, and patient satisfaction. A common feature of most of these outcome measures is that data are collected by questionnaires directly from patients, residents, employees, insured, or HMO-enrolled beneficiaries.

24.6.1.1 Patient-Based Measures of Health Status

Clinicians can make use of a variety of measures which are disease-specific, system- or organ-specific, function-specific (such as instruments that examine sleep or sexual function), or problem-specific (such as back pain) to explore the full range of patients' experience. Disease-specific health status measures have been developed for nearly all chronic conditions, including, for example, asthma, cancer sites, cardiovascular diseases, diabetes, rheumatoid arthritis, prostate disease, epilepsy, hypertension, pneumonia, and migraine (Guyatt et al. 1995). But if there is interest to go beyond the specific illness and to compare the impact of treatments on health-related quality of life (HRQL) across diseases or conditions, one will require a more comprehensive assessment. None of the disease-specific, system- or organ-specific, function-specific, or problem-specific measures are adequate for comparisons across conditions. These comparisons require generic measures designed for administration to people with any underlying health problem (or no problem at all) that cover all relevant areas of HRQL (Guyatt et al. 1995).

Generic health-status questionnaires are usually designed to establish separate scales including physical, mental, and social health, as suggested by the well-known definition of health by the WHO (1947). There are numerous generic health-status measures – for a review and description, see, e.g., Spilker (1995) or McDowell and Newell (1996). Three of these are very popular and have become standard in the health status field: The 36-Item Short Form Questionnaire (SF-36) (Ware and Sherbourne 1992), the Sickness Impact Profile (SIP) (Bergner et al. 1981), and the Quality of Well-Being Scale (QWB) (Kaplan and Anderson 1988). The psychometric properties of these instruments are sufficiently tested, and the reliability is considered high (Petitti 1998a).

In particular, the SF-36 (a shortened version of a battery of 149 health status questions) is one of the most widely accepted, extensively translated, and tested instruments around the world (Tseng et al. 2003). It satisfies rigorous psychometric criteria for validity and internal consistency. Clinical validity was shown by the distinctive profiles generated for each condition, each of which differed from that in the general population in a predictable manner. Furthermore, SF-36 scores were lower in referred patients than in patients not referred and were closely related to general practitioners' perceptions of severity (Garratt et al. 1993).

The SF-36 was designed for use in clinical practice and research, health policy evaluations, and general population surveys. It includes one multi-item scale that assesses eight health concepts: (1) limitations in physical activities because of health problems; (2) limitations in social activities because of physical or emotional problems; (3) limitations in usual role activities because of physical health problems; (4) bodily pain; (5) general mental health (psychological distress and well-being); (6) limitations in usual role activities because of emotional problems; (7) vitality (energy and fatigue); and (8) general health perceptions. See also the measurement concept in Fig. 24.3 and an excerpt of the questionnaire in Fig. 24.4. The survey was constructed for self-administration by persons 14 years of age and older and for administration by a trained interviewer in person or by telephone (Ware and Sherbourne 1992).

In the late 1980s, a European group of researchers started to develop a generic health-status measure – the European Quality of Life Scale (EQ-5D) – simultaneously in several European languages (EuroQol Group 1990; Brooks 1996). The EuroQol Group consisted originally of a network of international multidisciplinary researchers from Europe but nowadays includes members from Canada, Japan, New Zealand, Singapore, South Africa, and the USA. The EQ-5D self-report questionnaire comprises five dimensions of health (mobility, self-care, usual activities, pain/discomfort, anxiety/depression) rated on three levels (no problems, some/moderate problems, extreme problems). A unique EQ-5D health state is defined by combination of these dimensions. EQ-5D is a public domain instrument (http://www.euroqol.org/index.htm).

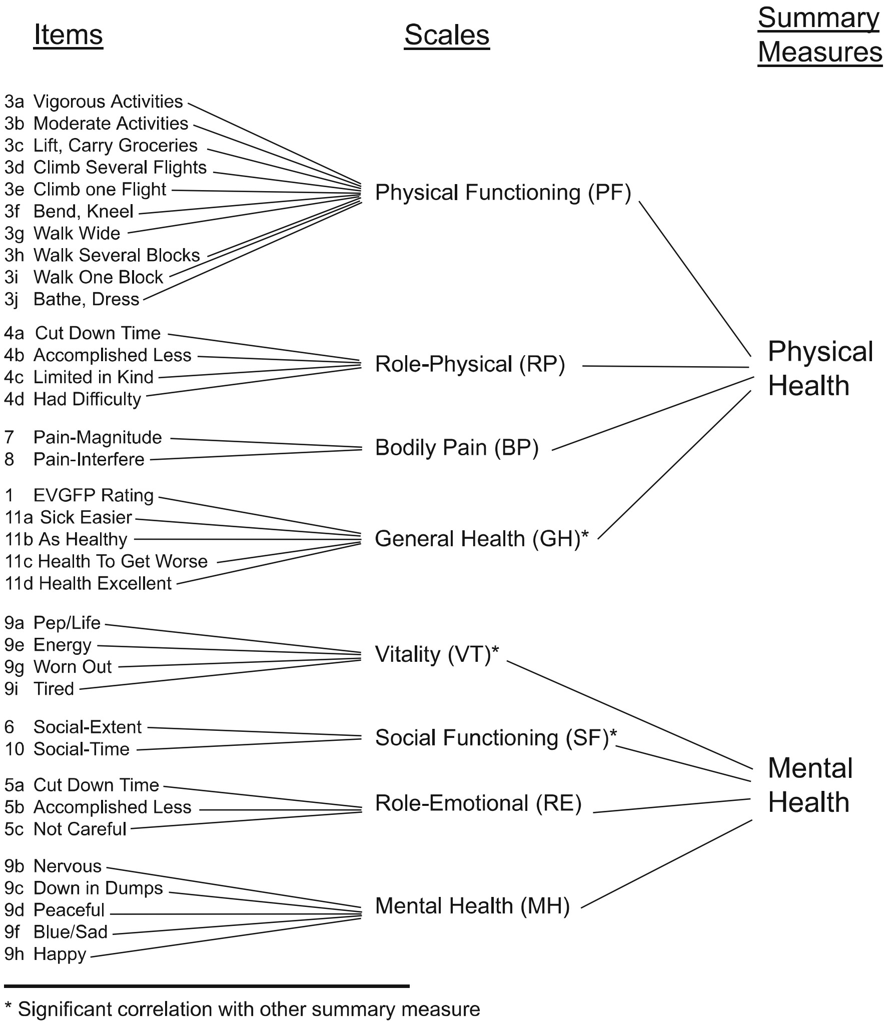
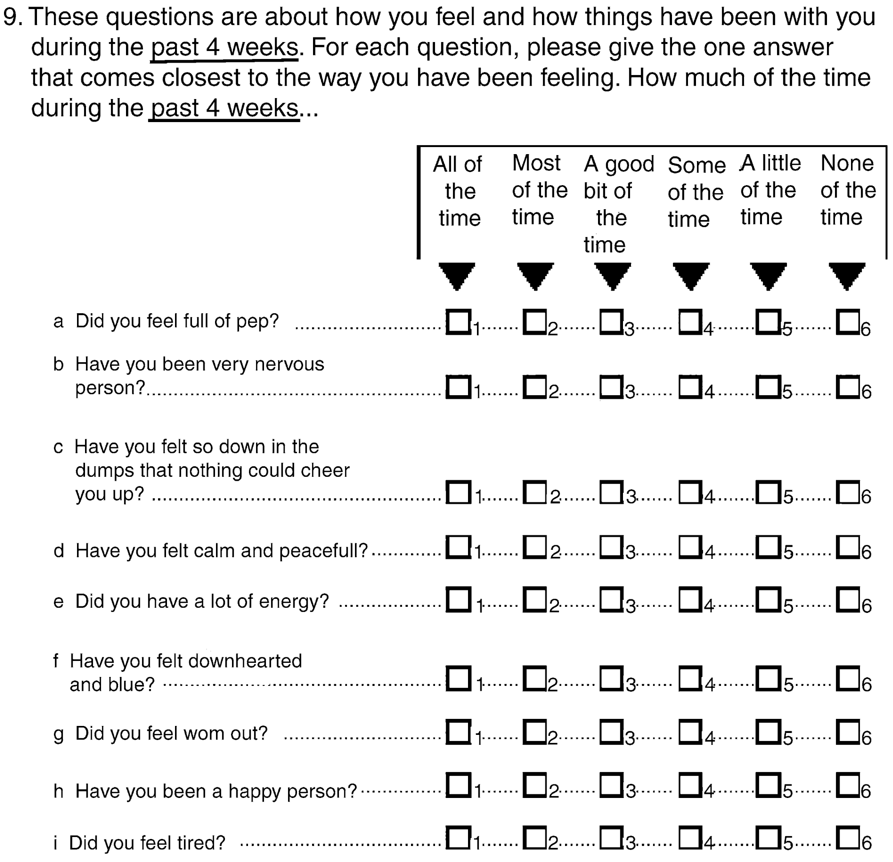


Fig. 24.3 The SF-36 measurement concept (Source: SF-36 Psychometric Considerations (http://www.sf-36.org/tools/sf36.shtml))

24.6.1.2 Adjusted Life Expectancy

Life expectancy, even without any adjustments, is already a rather complex measure. It is defined as the average future lifetime of a person at birth and is calculated from a current life table (the key tool of actuaries for some 200 years). Consider a large group, or “cohort,” of persons, who were born on the same day. If an actuary could follow the cohort from birth until death, he or she could record the number of individuals alive at each birthday – age x, say – and the number dying during



9. These questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the past 4 weeks...

	All of the time	Most of the time	A good bit of the time	Some of the time	A little of the time	None of the time
	▼	▼	▼	▼	▼	▼
a Did you feel full of pep?	□1	□2	□3	□4	□5	□6
b Have you been very nervous person?	□1	□2	□3	□4	□5	□6
c Have you felt so down in the dumps that nothing could cheer you up?	□1	□2	□3	□4	□5	□6
d Have you felt calm and peacefull?	□1	□2	□3	□4	□5	□6
e Did you have a lot of energy?	□1	□2	□3	□4	□5	□6
f Have you felt downhearted and blue?	□1	□2	□3	□4	□5	□6
g Did you feel wom out?	□1	□2	□3	□4	□5	□6
h Have you been a happy person?	□1	□2	□3	□4	□5	□6
i Did you feel tired?	□1	□2	□3	□4	□5	□6

Fig. 24.4 Excerpt from the SF-36 questionnaire: Item 9 (Source: SF-36 Health Survey Scoring Demonstration (http://www.sf-36.org/demos/SF-36.html))

the following year. The ratio of these is the probability of dying at age x, usually denoted by $q(x)$. It turns out that once the $q(x)$'s are all known, the life table is completely determined. In practice, such "cohort life tables" are rarely used, in part because individuals would have to be followed for up to 100 years, and the resulting life table would reflect historical conditions that may no longer have relevance. Instead, one generally works with a period, or current, life table. This summarizes the mortality experience of persons of all ages in a short period, typically 1 year or 3 years. More precisely, the death probabilities $q(x)$ for every age x are computed for that short period, often using census information gathered at regular intervals (e.g., every 10 years in the US). These $q(x)$'s are then applied to a hypothetical cohort of 100,000 people over their life span to create a current life table (Strauss and Shavelle 2000).

Several approaches have been developed to adjust life expectancy for aspects of health-related quality of life (Drummond et al. 1997). Most often used are the concepts of quality-adjusted life years (QALY) on the one hand and the concept of disability-adjusted life years (DALY) on the other hand (cf. chapter ▶Descriptive Studies of this handbook).

A quality-adjusted life year is a measure that assigns a (utility) value, often called Q, between 0 and 1 to each health state of a year with 0 representing death and 1 representing perfect health. The Q factors are then multiplied with the time spent in the corresponding health states, and these weighted times finally are summed up to achieve the QALY.

Three methods are used alternatively to establish a set of consistent Q values, all derived from consumer choice theory, which describes how consumers decide what to buy on the basis of two fundamental elements: their budget constraints and their preferences. Consumer preferences for different consumables are often represented by the concept of "utility" (Torrance et al. 1972; Torrance 1987; Mankiw 1998). The techniques proposed to measure the utility of specific health states on a linear scale were the Von Neumann-Morgenstern "standard gamble," the "time trade-off" method, and direct scaling techniques (e.g., category rating). These were claimed to produce equivalent and reliable results, but the time trade-off is easier to administer than each of the other two techniques (O'Connor 1993). The results of a simultaneous test of all three methods were that subjects found the time trade-off task the easiest, the standard gamble slightly more difficult (but probably impossible without some props), and the direct scaling task the most difficult. Only the time trade-off task was considered to be capable of being executed without a well-trained interviewer (Torrance 1976; O'Connor 1993). Nevertheless, direct scaling methods are commonly used to derive preferences (Petitti 1998a), probably without observing the necessary methodological diligence.

In a standard gamble, the rater (that is, the person to establish the utilities) must choose between two alternatives. One alternative has a certain outcome (that is, the health state to be rated) and the other involves a gamble with two possible outcomes: the best health state (usually complete health), which is described as occurring with a probability, p, or an alternative state, the worst state (usually death) which is described as occurring with probability 1-p. The probability p is varied until the rater is indifferent to the alternative which is certain and the gamble that may bring the better health state. The time trade-off task also entails a choice between two alternatives, but neither is a gamble. Each is a different health state but for differing periods of time. The rater is asked to value a choice of being in a less desirable health state for a longer time followed by death compared with being in a more desirable state for shorter period of time followed by death. The time in the less desirable health state then is decreased to the point of indifference. In category rating, raters sort the health states into a specified number of categories, and equal changes in preference between adjacent categories are assumed to exist (Petitti 1998a).

QALYs have been widely criticized on ethical, conceptual, operational, and methodological grounds. To begin with the last ones, Prieto and Sacristán (2003) have recently pointed to a considerable problem, which results from the numerical nature of its constituent parts. The appropriateness of the QALY arithmetical operation is compromised by the essence of the utility scale: while life years are expressed in a ratio scale with a true zero, the utility is an interval scale where 0 is an arbitrary value for death. In order to be able to obtain coherent results, both scales would have to be expressed in the same units of measurement. The different nature of these two factors jeopardizes the meaning and interpretation of QALYs. By a simple general linear transformation of the utility scale, the authors demonstrate that the results of the multiplication are not invariant and offer a mathematically solution to these limitations through an alternative calculation of QALYs by means of operations with complex numbers so that the new QALYs have a real part (length of life) and an imaginary part (utility). The revisited formulation of the QALYs provides a less dramatic adjustment of years of life than that implied by the multiplicative model. The maximum penalization represented by living in a suboptimal state of health is capped at 30% of the total time lived in that state, in contrast to the case of the multiplicative model, where the penalization can reach 100% (Prieto and Sacristán 2003).

Ethical concerns arise when QALYs are used in cost-effectiveness or cost-utility analysis for evaluation of alternative health policies, treatment programs, or setting of priorities. A simple ratio cost/QALY is commonly calculated in this type of analysis in order to compare cost-effectiveness of treatments, intervention and programs, etc. But it has been pointed out, among other arguments, that investing in the interventions that have the lowest-cost per QALY ignores the principle of equity (Drummond 1987). In addition, QALYs share a problem of life expectancy as a measure of outcome: they discriminate against the aged and the disabled because these groups of persons have fewer life years to gain from an intervention (Harris 1987).

The main other type of commonly used summary measure which combines information on mortality and morbidity is the disability-adjusted life year. The DALY is the best known example of a “health gap” summary measure, which quantifies the gap between a population’s actual health and a defined goal used to quantify the burden of disease in a country, region, or on the global level (Murray and Lopez 1996). However, DALYs share most of the methodological and ethical difficulties with QALYs, such as utility-weighting and discounting health benefits. The discrimination of elderly people is even more pronounced than with QALYs as an additional age-weighting is performed when constructing DALYs, whereby years lost during the productive phase of life get a higher weight than years lost in childhood or at a more advanced age (Gericke and Busse 2003). Related concepts are disability-free life years (Sullivan 1971) and healthy life expectancy (Robine and Ritchie 1991) which may be based on surveys. DALYs can be calculated exclusively based on life tables from census data and cross-sectional data from official disability statistics (if necessary, on a sample base). The so-called Sullivan method to adjust the conventional life table for disability consists of applying disability rates calculated from cross-sectional data to the person-years of the conventional life table. This calculation results into new estimates of the person-years lived in disability, and the complement of the later, the person-years lived free of disability, the DALYs (Guend et al. 2002).

A related measure is disability-adjusted life expectancy (DALE) used by WHO in a controversial report to display the burden of disease by cause, sex, and mortality stratum in WHO regions (WHO 2000). The disability rates used in the WHO calculations relied on subjective and expert assessment and not on empirical data due to data limitations in many nations.

24.6.1.3 Patient Satisfaction

Interest in measuring satisfaction with health care has grown considerably in recent years around the world, and there is a large and expanding literature in this field. Patient satisfaction and its measurement are undoubtedly important issues for public policy analysts, health-care managers, practitioners, and users. Nevertheless, measurement of satisfaction often lacks a clear definition. In particular, it is not always well understood by the people who measure it that satisfaction is a relative concept which can be measured only against individuals' expectations, needs, or desires (Wüthrich-Schneider 2000; Crow et al. 2002). Despite problems with establishing a tangible definition of "satisfaction" and difficulties with its measurement (among other things those which are predicted by the well-known theory of cognitive dissonance, cf. Festinger 1957), the concept continues to be widely used. However, in many instances, when investigators claim to be measuring satisfaction, more general evaluations of health-care services are being undertaken that tend to result in high levels of satisfaction being recorded (Crow et al. 2002).

Historically patient satisfaction surveys have focused on inpatient health-care services, but in recent years, investigations of patient satisfaction have been carried out in outpatient settings as well. In Germany, for example, a recently developed questionnaire to measure patient satisfaction in generalist and specialist ambulatory medical care comprises 27 single items divided into the four dimensions: "professional competence," "physician-patient interaction," "information," and "organization of the practice." This concept has been tested in a survey of 3,487 patients in 123 physician practices (Gericke et al. 2004). A former international study of patients' priorities with respect to general practice care collected data by postal surveys in UK, Norway, Sweden, Denmark, the Netherlands, Germany, Portugal, and Israel. The study results show that patients in different cultures and health-care systems have many views in common, particularly concerning doctor-patient communication and accessibility of services (Grol et al. 1999).

24.6.2 Assessing Efficiency of Care

In addition to measuring the output of health care in terms of healthy life gained, efficiency is another important dimension in assessing health service output. Unfortunately, the word efficiency is often used inappropriately to describe productivity, that is, relating episodes of care or number of procedures to the inputs or costs (Gray 1997). Efficiency refers to the health system's ability to use whatever resources it has to maximum effect (Le Grand 1998). Efficiency has three levels: technical, productive, and allocative efficiency. Technical efficiency answers the narrow question of whether the same or a better outcome could be obtained by using less of one type of input (Palmer and Torgerson 1999). It is based on effectiveness. Productive or internal efficiency is achieved when the maximum possible improvement in outcome is obtained from a given level of resource inputs or when costs are minimized to obtain a given level of output (Donaldson and Gerard 1993; Palmer and Torgerson 1999). Prerequisite for productive efficiency is technical efficiency.

Allocative or external efficiency refers to the way resources are divided between alternative uses within the health sector (Barr 1998). It implies productive efficiency (Donaldson and Gerard 1993). The theoretical foundation of allocative efficiency rests on the Pareto criterion: a resource allocation is efficient if it is impossible to move to an alternative allocation which would make some people better off and nobody worse off (Begg et al. 1997b). Among other conceptual difficulties, strict adherence to this principle would preclude changes that would make many people much better off at the expense of a few made slightly worse off (Palmer and Torgerson 1999). Therefore, an operational utilitarian decision rule is often used instead: allocative efficiency is achieved when resource allocation maximizes social welfare (Palmer and Torgerson 1999). Cost-effectiveness studies as a tool to put the concept of operational efficiency in health care into practice have already been summarized in Sect. 24.4.1. Cost-benefit studies can address questions of allocative efficiency comparing interventions between different sectors, as output of care is measured in monetary units. As this is politically and ethically difficult to accept for many non-economists, cost-benefit analyses of health interventions are seldom performed.

24.6.3 Assessing the Outcome of Health Systems

In principle, the same methods are used to assess the outcome of health systems which are used to assess the outcome of health services within a country. However, problems with data quality, definitions, and comparability across different cultures make comparisons between different health systems more difficult than health service research limited to a particular country (Schwartz and Busse 2003). As decision makers in countries of all levels of development are faced with common problems as they struggle to make appropriate choices to improve the performance of their health systems, the interest of politicians and scientists in comparative health system research has grown rapidly during the last two decades. A common goal of researchers is to provide policy decision makers and managers with the best available evidence in order to inform policy decision making. In analogy to evidence-based medicine, this movement has been termed evidence-based health policy or evidence-based health care. However, the evidence-base on how to improve the performance of health systems is still weak (Murray and Evans 2003).

24.6.3.1 Cross-Sectional Comparisons

Two methodological approaches are commonly used in comparative health system research: a cross-sectional approach comparing a number of parameters at a particular point in time and a longitudinal approach comparing the development of parameters over a defined time period. To illustrate the advantages and disadvantages of both approaches, we will focus here on two examples. The first is a summary of the approach taken by the World Health Organization (WHO) to assess health system performance on a global scale. In 1998, WHO embarked on a project to assess the health system performance of its member states, culminating in the World Health Report 2000, in which countries' health systems were ranked according to their performance. Health system performance was measured according to the level and distribution of population health, responsiveness, and fairness in financing (World Health Organization 2000; Murray and Evans 2003). Although the provision of comparative data on health system characteristics is recognized as important in improving health-care systems, the report has elicited heavy criticism, summarized by Gravelle et al. (2003). These included the purpose of the exercise (Williams 2001), the definition of some of the performance measures (Braveman et al. 2001), the quality of data (McKee 2001; Williams 2001), and mixed messages (Navarro 2000). Gravelle et al. (2003) furthermore demonstrated that the efficiency rankings and estimates of the magnitude of inefficiency in countries were not robust when compared with other, no less reasonable, methodological choices concerning the econometric methods used. The final rankings for a number of EU countries and ranking results concerning patient satisfaction with health systems are illustrated in Table 24.4 and compared to a number of other parameters which are commonly used to measure the input, process, and outcome of a health system.

It can be noted that parameters differ widely between countries at a similar level of national income and development. Some factors show a close correlation, for example, the health score with patient satisfaction or hospital bed provision with hospital utilization. On the other hand, satisfaction with the health system does not correlate at all with the overall WHO ranking of the health system performance. This results in contradictory results for countries like Denmark and Finland on the one hand, and Spain on the other (Schwartz and Busse 2003). Table 24.4 illustrates some of the issues surrounding the interpretation of cross-sectional data. Different data sources can vary substantially on the same measure. Such discrepancies – if detected at all – demand a thorough investigation of possible causes. A common cause are differences in the numerator, for example, differences between licensed and practicing doctors or beds in acute care hospitals or in all hospitals. Differences in the denominator are also important. For instance, for the measurement of neonatal mortality, it makes a difference whether all births on the territory of a country are counted or all births of nationals of that country (Schwartz and Busse 2003).

The most important difficulty with cross-sectional comparisons of health systems from a policy perspective is probably that health output measured in terms of reduced mortality and health system performance are correlated in a contradictory way. If a country reacts in an appropriate way to high mortality rates and invests in

Table 24.4 Selected input, process, and outcome parameters for some European countries, around 1997. Data from the OECD Health Dataset 2001, the WHO Health for All database 2003, and the World Health Report 2000 (World Health Organization 2000) (Adapted from Schwartz and Busse (2003))

	Financial input: % of GDP (1998)	Structure: Hospital beds/1,000 population (1997)	Structure: Doctors/ 1,000 population (1997)	Process: Hospital cases/100 pop./year (1996)	Process: Hospital days/capita (1996)	Process: Ambulatory doctor-patient contacts/year (1996)	Outcome: Neonatal mortality/ 1,000 (1998)	Outcome: Satisfaction with health system in % [Ranking within EU] (1998)	Outcome: Overall ranking of health system within EU (1999)
Austria	8.0	9.1	2.9	25.1	2.6	6.3	4.9	72.7 [3]	4
Belgium	8.6	7.3	3.7	20.0	2.2	8.0	5.6	62.8 [7]	11
Denmark	8.3	4.6	3.3	19.8	1.4	5.7 (6.6)[a]	4.7	90.6 [1]	15
Finland	6.9	7.9	3.0	26.9	3.2	4.3	4.1	81.3 [2]	14
France	9.4	8.6	3.0	22.5	2.6	6.5	4.6	65.0 [6]	1
Germany	10.3	9.4	3.4	19.7	2.8	6.5	4.7	57.5 [9]	13
Greece	8.4	5.0	4.1	–	1.2	–	5.7 (6.7)[a]	15.5 [15]	6
Ireland	6.8	–	2.1	15.1	1.1	–	6.2	57.9 [8]	10
Italy	8.2	5.8	5.8	18.5	1.7	–	5.3	20.1 [13]	2
Luxembourg	6.0	8.1	3.0 (2.4)[a]	–	2.8	2.9	5.0	66.6 [5]	7
Netherlands	8.7	11.3 (5.3)[a]	–	11.1	3.6	5.4	5.0	69.8 [4]	8
Portugal	7.7	4.1	3.1	11.4	1.1	3.2	5.9	16.4 [14]	5
Spain	7.0	–	2.9	11.4 (10.0)[a]	1.1	–	5.7 (4.9)[a]	43.1 [12]	3
Sweden	7.9	4.0 (5.2)[a]	3.1	18.1	1.3	2.9	3.5	57.5 [9]	12
United Kingdom	6.8	4.4	1.7	15.0 (23.1)[a]	1.3	6.1	5.7	57.0 [11]	9

[a] More than 10% difference between OECD and WHO datasets

health system infrastructure, mortality would fall as a result assuming effectiveness of the measures taken. This longitudinal result cannot be measured in cross-sectional studies. Therefore cross-sectional comparisons cannot indicate whether a high level of inputs in a particular country has obviated even higher mortality rates and we only see average mortality in this country or whether there truly exists an inefficient input-output relation.

24.6.3.2 Longitudinal Comparisons

The other approach consists in comparing the development of input, process, and output parameters in different health systems in a longitudinal perspective. In the 1980s, time series analyses on “avoidable mortality” marked the first attempts at international longitudinal comparisons (Bunker et al. 1994; Charlton and Velez 1986). A common measure for comparing health systems in a longitudinal way is life expectancy. In Fig. 24.5, the development of life expectancy at birth is depicted for a number of selected European countries compared to the EU average for the time period 1970–2000.

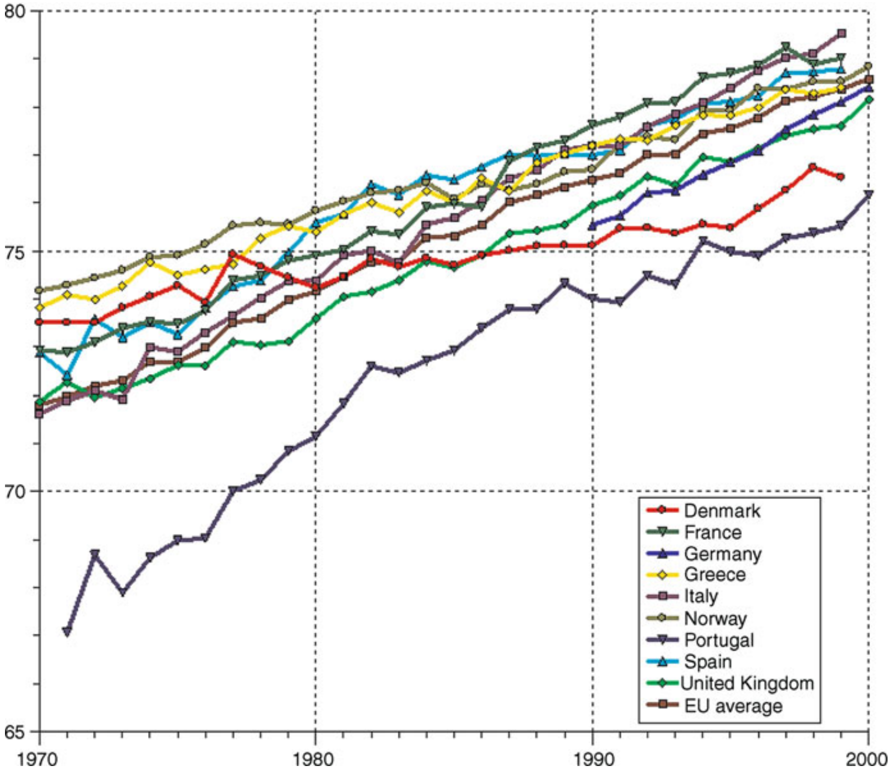


Fig. 24.5 Life expectancy at birth in selected member countries of the European Union 1970–2000. Calculated with data from WHO Health for All database 2003

This is often done although it is well known that life expectancy is influenced by many variables outside the scope of the health system, such as the level of socioeconomic development. However, in general, mortality (on which the calculation of life expectancy is based) is an output measure which is relatively insensitive to common health service endeavours (Schwartz and Busse 2003):

- The overwhelming part of mortality is not amenable to health service activity ("avoidable mortality") but natural mortality.
- In particular for men, a substantial proportion of deaths is due to traffic accidents.
- The commonplace argument that mortality figures do not respond quickly to changes had to be revised after the experience in Russia after the breakdown of the Soviet Union, where life expectancy at birth for males fell by approximately 6 years between 1990 and 1994. Life expectancy in the Eastern part of Germany, however, increased substantially in the 1990s.

Relative changes over time are of particular importance for evaluation and policy decision making. This is illustrated in the development of life expectancy at birth in Denmark and Portugal. Whereas both countries had an average life expectancy at birth of 76 years in the year 2000, Portugal has massively improved on this measure since 1970, up from 67 years. Although life expectancy in Denmark has nominally also increased since 1970, up from 74 years, it has had the smallest relative increase in Western Europe – which is in fact a rather negative development and not an improvement.

24.7 Conclusions

As demonstrated in the examples discussed above, the combination of simple inputs and outputs can be of particular political importance, despite all the methodological difficulties and caveats. The fact that even if life expectancy were a good indicator of health production in the health-care system, the question of why a good result has occurred, that is, examining structure and process, would not have been answered. There is little consensus on how international comparisons of structures and processes should be performed. How inappropriate simplification of health system comparisons can be misleading is demonstrated by the "state versus free market" debate in Germany. Financing of German hospital care on the basis of *per diem* payments has been coined as inefficient, as this payment mechanism creates an incentive for longer hospital stays. Some economists have compared the German system with the US system, where hospital stays are usually shorter, claiming that this was due to payments according to diagnostic-related groups (DRGs). However, they did not consider that at that time, only hospital services for 15% of the population covered under the Medicare scheme were remunerated according to DRGs and that hospital costs per case in the USA were about twice as high as in Germany, "despite" the DRGs. Likewise, the expected rise in ambulatory care costs to compensate for early hospital discharge was not considered (Schwartz and Busse 2003).

International comparisons of health system outcomes along one-dimensional hypotheses have thus to be treated with great caution, in particular, because they are easily misunderstood by policy decision makers (Schwartz and Busse 2003).

References

Abraham S (1986) Analysis of data from a complex sample: the Health Examination Surveys. Am J Clin Nutr 43:839–843

ACRA (1999) A brief history of resource allocation in the NHS, 1948–98. Advisory Committee on Resource Allocation, Department of Health, London

Aday LA, Begley CE, Lairson DR, Slater CH (1998) Evaluating the healthcare system. Effectiveness, efficiency, and equity. Health Administration Press, Chicago

Adler GS (1994) A profile of the Medicare Current Beneficiary Survey. Health Care Financ Rev 15(4):153–163

American Child Health Association (1934) Physical defects: the pathway to correction. ACHA, New York

Armenian HK (1998) Case-control methods. In: Amenian HK, Shapiro S (eds) Epidemiology and health services. Oxford University Press, New York/Oxford, pp 135–155

Ash A, Porell F, Gruenberg L, Sawitz E, Beiser A (1989) Adjusting Medicare capitation payments using prior hospitalization data. Health Care Financ Rev 10(4):17–29

Asthana S, Gibson A, Moon G, Dicker J, Brigham P (2004) The pursuit of equity in NHS resource allocation: should morbidity replace utilisation as the basis for setting health care capitations? Soc Sci Med 58:539–551

Barr N (1998) The economics of the welfare state. Oxford University Press, Oxford

Bedford Z, Kafetz A (2009) Editors' letter. In: The Dr Foster Hospital Guide 2009. Dr Foster Research. London

Begg D, Fischer S, Dornbusch R (1997a) Demand, supply, and the market. In: Economics. McGraw Hill, London, pp 30–43

Begg D, Fischer S, Dornbusch R (1997b) Introduction to welfare economics. In: Economics. McGraw Hill, London, pp 240–259

Bellach B-M, Knopf H, Thefeld W (1998) Der Bundesgesundheitssurvey 1997/98. Gesundheitswesen 60(Suppl 2):59–68

Bennett AC (1978) Improving management performance in health care institutions. American Hospital Association, Chicago

Bergmann E, Kamtsiuris E (1999) Inanspruchnahme medizinischer Leistungen. Gesundheitswesen 61(Spec Issue):138–144

Bergner M, Bobbitt RA, Carter WB, Gilson BS (1981) The sickness impact profile: development and final revision of a health status measure. Med Care 19:787–805

Berrington de Gonzales A, Darby S (1994) Risk of cancer from diagnostic X-rays: estimates for the UK and 14 other countries. Lancet 363:345–351

Bitzer EM, Neusser S, Lorenz C, Dörning H, Schäfer T (2007) Krankenhaus-Rangfolgen nach Ergebnisqualität in der Hüftendoprothetik - Routinedaten mit oder ohne Patientenbefragungen? - Teil 2: Patientenbefragung in Kombination mit Routinedaten. GMS Med Inf Biom Epidemiol 3(1):Doc07

Bitzer EM, Grobe T, Dörning H, Schwartz FW (2010) BARMER GEK Report Krankenhaus 2010. Asgard, St. Augustin

Black N (1997) Health services research: saviour or chimera? Lancet 349:1834–1836

Black N, Langham S, Petticrew M (1995) Coronary revascularisation: why do rates vary geographically in the UK? J Epidemiol Community Health 49:408–412

Blumenthal D (1996) Quality of care: what is it? N Engl J Med 335:891–894

Brand DA, Newcomer LN, Freiburger A, Tian H (1995) Cardiologists' practices compared with practice guidelines: use of beta-blockade after acute myocardial infarction. J Am Coll Cardiol 26:1432–1436

Braveman P, Starfield B, Geiger HJ (2001) World Health Report 2000: how it removes equity from the agenda for public health monitoring and policy. BMJ 323:678–681

British Medical Association (2000) Clinical indicators (League tables) – a discussion paper. British Medical Association, London

Brook RH, Lohr KN (1985) Efficacy, effectiveness, variations, and quality. Boundary-crossing research. Med Care 23:710–722

Brook RH, Ware JE Jr, Rogers WH, Keeler EB, Davies AR, Donald CA, Goldberg GA, Lohr KN, Masthay PC, Newhouse JP (1983) Does free care improve adults' health? Results from a randomized controlled trial. N Engl J Med 309:1426–1434

Brook RH, Park RE, Chassin MR, Solomon DH, Keesey J, Kosecoff J (1990) Predicting the appropriate use of carotid endarterectomy, upper gastrointestinal endoscopy, and coronary angiography. N Engl J Med 323:1173–1177

Brook RH, McGlynn EA, Cleary PD (1996) Quality of health care. Part 2: Measuring quality of care. N Engl J Med 335:966–970

Brooks R (1996) EuroQol: the current state of play. Health Policy 37:53–72

Brückner G (1997) Developing a new system of health care statistics. A major challenge for all participants? Federal Statistical Office, Wiesbaden

Bunker JP, Frazier HS, Mosteller F (1994) Improving health: measuring effects of medical care. Milbank Q 72:225–258

Buring JE, Hennekens CH (1992) The women's health study: summary of the study design. J Mycardial Ischemia 4:27–29

Busse R (1995) Radiologie, Gesundheitsstrukturreform und Gesundheitssystemforschung – Stand, Entwicklungen und Herausforderungen. Akt Radiologie 5:127–130

Busse R, Wismar M (2002) Health target programmes and health care services – any link? A conceptual and comparative study (part 1). Health Policy 59:209–221

Cameron AC, Trivedi PK (1998) Regression analysis of count data. Cambridge University Press, Cambridge/New York/Melbourne

Campell DT, Stanley JC (1966) Experimental and quasi-experimental designs for research. Rand McNally, Chicago

Carr-Hill R (1989) Allocating resources to health care: RAWP (Resources Allocation Working Party) is dead-long live RAWP. Health Policy 13:135

Carr-Hill RA, Sheldon TA, Smith P, Martin S, Peacock S, Hardman G (1994) Allocating resources to health authorities: development of method for small area analysis of use of inpatient services. BMJ 309(6961):1046–1049

Charlton JR, Velez R (1986) Some international comparisons of mortality amenable to medical intervention. BMJ 292:295–301

Childs AW, Hunter ED (1972) Non-medical factors influencing use of diagnostic x-ray by physicians. Med Care 10:323–335

Cochran WG (1968) Sampling techniques. Wiley, New York

Cohen SB (1997) Sample design of the 1996 Medical Expenditure Panel Survey Household Component. MEPS Methodol Rep No 2. AHCPR Pub. No 97–0027, Rockville

Crow R, Gage H, Hampson S, Hart J, Kimber A, Storey L (2002) The measurement of satisfaction with healthcare: implications for practice from a systematic review of the literature. Health Technol Assess 6:32

Culyer A (1993) Health, health expenditures, and equity. In: VanDoorslaer E (ed) Equity in the finance and delivery of health care. Oxford University Press, Oxford

de Miguel JM (1971) A framework for the study of national health systems. Inquiry 12:10–24

Department of Health (2008) Resource allocation: weighted capitation formula, 6th edn. Department of Health, Leeds

Detsky AS, Naglie IG (1990) A clinician's guide to cost-effectiveness analysis. Ann Intern Med 113:147–154

Deutsche Röntgengesellschaft (2002) Pressemitteilung anlässlich des 83. Deutschen Röntgenkongresses vom 8. - 11. Mai 2002 in Wiesbaden

Diehr P, Yanez D, Ash A, Hornbrook M, Lin DY (1999) Methods for analyzing health care utilization and costs. Ann Rev Public Health 20:125–144

Donabedian A (1980) Explorations in quality assessment and monitoring. Health Administration Press, Ann Arbor, MI

Donabedian A (1985) The methods and findings of quality assessment and monitoring. Health Administration Press, Ann Arbor, MI

Donabedian A (1988) The quality of care: how can it be assessed? JAMA 260:1743–1748

Donaldson C, Gerard K (1993) Economics of health care financing: The visible hand. Macmillan, London

Doubilet P, Weinstein MC, McNeil JB (1986) Use and misuse of the term 'cost-effectiveness' in medicine. N Engl J Med 314:253–256

Dr Foster (ed) (2004a) Hospital guide – methodology. http://www.drfosterhealth.co.uk/hospital-guide/methodology/. Accessed 29 Mar 2011

Dr Foster (ed) (2004b) Consultant guide – methodology. http://www.drfosterhealth.co.uk/consultant-guide/methodology.aspx. Accessed 29 Mar 2011

Drummond MF (1987) Ressource allocation decisions in health care. A role for quality of life assessments? J Chronic Dis 40:605–616

Drummond MF, O'Brien B, Stoddart GL, Torrance GW (1997) Methods for the economic evaluation of health care programmes, 2nd edn. Oxford University Press, Oxford

Dunn DL, Rosenblatt A, Taira DA, Lattimer E, Bertko J, Stoiber T (1996) A comparative analysis of methods of health risk assessment. Society of Actuaries, Schaumburg, IL

Elinson J (1987) Advances in health assessment discussion panel. J Chronic Dis 40(Suppl 1): 83S–91S

Ellis R, Pope G, Iezzoni L, Ayanian J, Bates D, Burstin H, Ash A (1996) Diagnosis-based risk adjustment for Medicare capitation payments. Health Care Financ Rev 17(3):101–128

Ellwood PM (1988) Outcomes management: a technology of patient experience. N Engl J Med 318:1549–1556

Epstein RS, Sherwood LM (1996) From outcomes research to disease management. A guide for the perplexed. Ann Intern Med 124:832–837

Etchason J, Petz L, Keeler E, Calhoun L, Kleinman S, Snider C, Fink A, Brook R (1995) The cost effectiveness of preoperative autologous blood donations. N Engl J Med 332:719–724

EuroQol Group (1990) EuroQol – a new facility for the measurement of health-related quality of life. Health Policy 16:199–208. http://www.euroqol.org/. Accessed 29 Mar 2011

Fein R (1971) On measuring economic benefits of health programs. In: McLachlan G, McKeown T (eds) Medical history and medical care. Oxford University Press, London

Ferris G, Roderick P, Smithies A, George S, Gabbay J, Couper N, Chant A (1998) An epidemiological needs assessment of carotid endarterectomy in an English health region. Is the need being met? BMJ 317:447–451

Festinger L (1957) A theory of cognitive dissonance. Stanford University Press, Stanford, CA

Fowler FJ Jr, Barry MJ, Lu-Yao G, Roman A, Wasson J, Wennberg JE (1993) Patient-reported complications and follow-up treatment after radical prostatectomy. The national Medicare experience: 1988–1990 (updated June 1993). Urology 42:622–629

Freedman MA in collaboration with the CDC Health Status Indicators Consensus Work Group (1991) Health status indicators for the year 2000. Health People 2000 Statistical Notes, vol 1 (no. 1). National Center for Health Statistics, Hyattsville

Garratt AM, Ruta DA, Abdalla MI, Buckingham JK, Russell IT (1993) The SF36 health survey questionnaire: an outcome measure suitable for routine use within the NHS? BMJ 306: 1440–1444

Gericke C, Busse R (2003) Gesundheitsökonomische Aspekte der Pharmakotherapie älterer Menschen. Arzneimittelforschung Drug Res 53:918–921

Gericke CA, Schiffhorst G, Busse R, Häussler B (2004) Ein valides Instrument zur Messung der Patientenzufriedenheit in ambulanter haus- und fachärztlicher Behandlung: das QUALISKOPE-A. Das Gesundheitswesen 66:723–731

Gillings D, Makuc D, Siegel W (1981) Analysis of interrupted time series mortality trends; an example to evaluate regionalized perinatal care. Am J Public Health 71:38–46

Glaeske G, Schicktanz C (2010) BARMER GEK Arzneimittel-Report 2010. Asgard, St. Augustin

Gold MR, Siegel JE, Rusek KB, Weinstein MC (1996) Cost-effectiveness in health and medicine. Oxford University Press, New York

Gravelle H, Jacobs R, Jones AM, Street A (2003) Comparing the efficiency of national health systems: a sensitivity analysis of the WHO approach. Appl Health Econ Health Policy 2: 141–147

Gray JAM (1997) Assessing the outcomes found. In: Evidence-based healthcare. Churchill Livingstone, New York, pp 103–154

Greene WH (2003) Econometric Analysis, 5th edn. Prentice and Hall, Upper Saddle River

Grimshaw JM, Wilson B, Campbell M, Eccles M, Ramsay C (2001) Epidemiological methods. In: Fulop N, Allen P, Clarke A, Black N (eds) Studying the organisation and delivery of health services: research methods. Routledge, London, pp 56–72

Grobe T, Dörning H, Schwartz FW (2006) GEK-Report ambulant-ärztliche Versorgung 2006. Asgard, St. Augustin

Grobe T, Dörning H, Schwartz FW (2011) BARMER GEK Arztreport 2011. Asgard, St. Augustin

Grol R, Wensing M, Mainz J, Ferreira P, Hearnshaw H, Hjortdahl P, Olesen F, Ribacke M, Spenser T, Szecsenyi J (1999) Patients' priorities with respect to general practice care: an international comparison. Fam Pract 16:4–11

Guend H, Stone-Newsom R, Swallen K, Lasker A, Kindig D (2002) State disability adjusted life expectancy using census disability. University of Wisconsin, Madison

Guyatt GH, Naylor CD, Juniper E, Heyland DK, Jaeschke R, Cook RD for the Evidence Based Medicine Working Group (1995) How to use articles about health-related quality of life measurements. Centre for Health Evidence, Edmonton. http://www.cche.net/usersguides/life.asp. Accessed 28 May 2004

Hannan EL, O'Donnell JF, Kilburn H Jr, Bernard HR, Yazici A (1989) Investigation of the relationship between volume and mortality for surgical procedures performed in New York State hospitals. JAMA 262:503–510

Hansen MH, Hurwitz WN, Madow WG (1953) Sample survey methods and theory, vols I and II. Wiley, New York/London

Harris J (1987) QUALYfying the value of life. J Med Ethics 13:117–123

Heller G, Günster C (2008) Mit Routinedaten Qualität in der Medizin sichern. GGW 8:26–34

Herrin J, Etchason JA, Kahan JP, Brook RH, Ballard DJ (1997) Effect of panel composition on physician ratings of appropriateness of abdominal aortic aneurysm surgery: elucidating differences between multispecialty panel results and specialty society recommendations. Health Policy 42:67–81

Hill AB (1965) The environment and disease: association or causation? Proc R Soc Med 58: 295–300

Hofer TP, Hayward RA, Greenfield S, Wagner EH, Kaplan SH, Manning WG (1999) The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. JAMA 281:2098–2105

Iezzoni LI (1994) Risk and outcome. In: Iezzoni LI (ed) Risk adjustment for measuring healthcare outcomes. Health Administration Press, Ann Arbor, Michigan, pp 1–28

Ihle P, Köster I, Küpper-Nybelen J, Schubert I (2008) Experiences with a person-related and population-based sickness fund sample (1997–2007) for pharmacoepidemiological and health care utilization research. Abstract of the 24th International Conference on Pharmacoepidemiology & Therapeutic Risk Management. Copenhagen, Denmark, August 17–20, 2008 Pharmacoepidemiol Drug Saf 17(Suppl 1): S242

Ingber MJ (1998) The current state of risk adjustment technology for capitation. J Ambul Care Manag 21(4):1–28

Institute of Medicine (1994) Health services research: opportunities for an expanding field of inquiry – An interim statement. National Academies Press, Washington, DC

Jacobson B, Mindell J, McKee M (2003) Hospital mortality league tables. BMJ 326:777–778

Jencks SF, Williams DK, Kay TL (1988) Assessing hospital-associated deaths from discharge data. The role of length of stay and comorbidities. JAMA 260:2240–2246

Kahn HA, Sempos CT (1989) Statistical methods in epidemiology. Oxford University Press, New York/Oxford

Kaplan RM, Anderson JP (1988) A general health policy model: update and applications. Health Serv Res 23:203–205

Keeler EB, Kahn KL, Draper D, Sherwood MJ, Rubenstein LV, Reinisch EJ, Kosecoff J, Brook (1990) Changes in sickness at admission following the introduction of the prospective payment system. JAMA 264:1962–1968

Kelsey JL, Petitti DB, King AC (1998) Key methodologic concepts and issues. In: Brownson RC, Petitti DB (eds) Applied epidemiology. Oxford University Press, New York, Oxford, pp 35–69

Kendall MG, Stuart A (1958) Advanced theory of statistics, vol 1. Charles Griffin and Co., London

Kindig DA (1997) Different populations, different needs? In: Purchasing population health: paying for results. The University of Michigan Press, Ann Arbor, pp 133–148

Kish L (1965) Survey sampling. Wiley, New York

Kitagawa EM, Hauser PM (1973) Differential mortality in the United States – a study in socioeconomic epidemiology. Harvard University Press, Cambridge

Kjerulff KH, Erickson BA, Langenberg PW (1996) Chronic gynecological conditions reported by U.S. women: finding from the National Health Interview Survey 1984 to 1992. Am J Public Health 86:195–199

Kohn R, White KL (eds) (1976) Health care – an international study. Oxford University Press, Oxford/New York/Toronto

Krishnaiah PR, Rao CR (eds) (1994) Handbook of statistics 6 - Sampling, 2nd edn. Elsevier Science Publishers, Amsterdam

Kronick R, Dreyfus T, Lee L, Zhou Z (1996) Diagnostic risk adjustment for Medicaid: the disability payment system. Health Care Financ Rev 17(3):7–33

Kronick R, Gilmer T, Dreyfus T, Ganiats T (2002) CDPS-Medicare: the chronic illness and disability payment system modified to predict expenditures for Medicare beneficiaries. Final report to CMS. University of California, San Diego

Kurth B-M (2007) The German Health Interview and Examination Survey for Children and Adolescents (KiGGS): an overview of its planning, implementation and results taking into account aspects of quality management. In: KiGGS – Principal publication, methodology and conduct of field work. Bundesgesundheitsbl – Gesundheitsforsch – Gesundheitsschutz 50(5–6):533–546

Lamers LM (1999) Pharmacy costs groups: a risk-adjusted for capitation payments based on the use of prescribed drugs. Med Care 37(8):824–830

Last JM (2001) A dictionary of epidemiology, 4th edn. Oxford University Press, Oxford

Le Grand J (1998) Financing health care. In: Feachem Z, Hensher M, Rose L (eds) Implementing health sector reform in Central Asia. World Bank, Washington, DC, pp 75–85

Leape LL (1994) Error in medicine. JAMA 272:1851–1857

Levy PS, Lemesshow S (1991) Sampling of populations: methods and applications. Wiley, New York

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22

Lipid Research Clinics Program (1984) The lipid research clinics coronary primary prevention trial results I. Reduction in incidence of coronary heart disease. JAMA 251:351–364

Lohr KN, Yordi KD, Their SO (1988) Current issues in quality of care. Health Aff 7:5–18

Luft HS, Bunker JP, Enthoven AC (1979) Should operations be regionalized? The empirical relation between surgical volume and mortality. N Engl J Med 301:1364–1369

Mankiw NG (1998) Principles of economics Harcourt. Brace, Boston

Mays N (1995) Geographical resource allocation in the English National Health Service, 1974–1994: the tension between normative and empirical approaches. Int J Epidemiol 24: 96–102

McCullagh P, Nelder JA (1983) Generalized linear models. Chapman and Hall, London/New York

McDowell I, Newell C (1996) Measuring health: a guide to rating scales and questionnaires, 2nd edn. Oxford University Press, New York/Oxford

McGlynn EA (1998) The outcomes utility index: Will outcomes data tell us what we want to know? Int J Qual Health Care 10:485–490

McKee M (2001) Measuring the efficiency of health systems. The world health report sets the agenda, but there's still a long way to go. BMJ 323:295–296

McPake B, Kumaranayake L, Normand C (2002) The demand for health and health services. In: Health economics. An international perspective. Routledge, London/New York, pp 12–19

Medicare Payment Advisory Commission (1998) Report to the Congress: context for a changing Medicare program. Medicare Payment Advisory Commission, Washington

Meyer N, Fischer R, Weitkunat R, Crispin A, Schotten K, Bellach B-M, Überla K (2002) Evaluation des Gesundheitsmonitorings in Bayern mit computer-assistierten Telefoninterviews (CATI) durch den Vergleich mit dem Bundesgesundheitssurvey 1998 des Robert Koch-Instituts. Gesundheitswesen 64:329–335

Murray CJL, Lopez AD (1996) The global burden of disease. A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020. Harvard University Press, Cambridge, MA

Murray JL, Evans DB (2003) Health systems performance assessment: goals, framework and overview. In: Murray JL, Evans DB (eds) Health systems performance assessment: debates, methods and empiricism. World Health Organization, Geneva, pp 3–20

Navarro V (2000) Assessment of the world health report 2000. Lancet 356:1598–1601

Nelder JA, Wedderburn EWM (1972) Generalized linear models. J R Stat Soc A 135:370–384

Neter J, Wasserman W (1974) Applied linear statistical models. Richard D. Irwin Inc., Homewood, IL

Neuhauser D, Lewicki AM (1975) What do we gain from the sixth stool guaiac? N Engl J Med 293: 226–228

Newhouse JP (1974) A design for a health insurance experiment. Inquiry 11:5–27

Newhouse JP (1986) Rate adjuster for Medicare under capitation. Health Care Financ Rev (Spec No):45–55

Newhouse JP (1993) Free for all? Lessons from the RAND health insurance experiment. Harvard University Press, Cambridge, MA

Newhouse JP, McClellan M (1998) Econometrics in outcomes research: the use of instrumental variables. Ann Rev Public Health 19:17–34

Newhouse JP, Manning WG, Morris CN, Orr LL, Duan N, Keeler EB, Leibowitz A, Marquis KH, Marquis MS, Phelps CE, Brook RH (1981) Some interim results from a controlled trial of cost sharing in health insurance. N Engl J Med 305:1501–1507

Newhouse JP, Manning WG, Keeler EB, Sloss EM (1989) Adjusting capitation rates using objective health measurers and prior utilization. Health Care Financ Rev 10(3):41–54

Newhouse JP, Buntin MB, Chapman JD (1997) Risk adjustment and Medicare: taking a closer look. Health Affairs 16(5):26–43

O'Connor R (1993) Issues in the measurement of health-related quality of life. Working paper 30. NHMRC National Centre for Health Program Evaluation Melbourne, Australia

OECD (1993) System of national accounts. OECD, Paris

OECD (2000) A system of health accounts, Version 1.0. OECD, Paris

Palmer RH (1997) Process-based measures of quality: the need for detailed clinical data in large health care databases. Ann Intern Med 127:733–738

Palmer S, Torgerson DJ (1999) Definitions of efficiency. BMJ 318:1136

Petitti DB (1998a) Epidemiological issues in outcomes research. In: Brownson RC, Petitti DB (eds) Applied epidemiology. Oxford University Press, New York/Oxford, pp 249–275

Petitti DB (1998b) Economic evaluation. In: Brownson RC, Petitti DB (eds) Applied epidemiology. Oxford University Press, New York/Oxford, pp 277–298

Petitti DB, Amster A (1998) Measuring the quality of health care. In: Brownson RC, Petitti DB (eds) Applied epidemiology. Oxford University Press, New York/Oxford, pp 299–321

Petitti DB, Sidney S (1989) Hip fracture in women: incidence, in-hospital mortality, and five-year survival probabilities in members of prepaid health plan. Clin Orthop 246:150–155

Phibbs CS, Bronstein JM, Buxton E, Phibbs RH (1996) The effects of patient volume and level of care at the hospital of birth on neonatal mortality. JAMA 276:1054–1059

Pigeot I, Ahrens W (2008) Establishment of a pharmacoepidemiological database in Germany: methodological potential, scientific value and practical limitations. Pharmacoepidemiol Drug Saf 17:215–223

Pope GC, Ellis RP, Ash AS, Liu C-F, Ayanian JZ, Bates DW, Burstin H, Iezzoni LI, Ingber MJ (2000) Principal inpatient diagnostic cost group model for Medicare risk adjustment. Health Care Financ Rev 21(3):93–118

Pope GC, Kautter J, Ellis RP, Ash AS, Ayanian JZ, Iezzoni LI, Ingber JM, Levy JM, Robst J (2004) Risk adjustment of Medicare capitation payments using the CMS-HCC model. Health Care Financ Rev 25(4):119–141

Prieto L, Sacristán JA (2003) Problems and solutions in calculating quality-adjusted life years (QALYs). Health Qual Life Outcome 1:80. http://www.hqlo.com/content/1/1/80. Accessed 30 July 2013

Rice N, Smith P (1999) Approaches to capitation and risk adjustment in health care: an international survey. ACRA paper 09, Department of Health, London

Rich MW, Shah AS, Vinson JM, Freedland KE, Kuru T, Sperry JC (1996) Iatrogenic congestive heart failure in older adults: clinical course and prognosis. J Am Geriatr Soc 44:638–643

Robert Koch Institute (2008a) German Health Interview and Examination Survey for Adults (DEGS). http://www.rki.de/cln_160/nn_217400/EN/Content/Health_Reporting/HealthlSurveys/Degs/degs_node.html?_nnn=true. Accessed 28 Mar, 2011

Robert Koch Institute (2008b) GEDA: telephone health survey 2008/2009. http://www.rki.de/cln_160/nn_675636/EN/Content/Health_Reporting/HealthlSurveys/Geda/ Geda_ node.html?_ nnn = true. Accessed 28 Mar 2011

Robert Koch Institute (2010) KiGGS - The German Health Interview and Examination Survey for Children and Adolescents ; Continuation of the KiGGS Study - wave 1 (2009–2012). http://www.kiggs.de/service/english/index.html. Accessed 28 Mar 2011

Robine JM, Ritchie K (1991) Healthy life expectancy: evaluation of global indicator of change in population health. BMJ 302:457–460

Rothgang H, Iwansky S, Müller R, Sauer S, Unger R (2010) BARMER GEK Pflegereport 2010. Asgard, St. Augustin

RTI International (2009) Methodology: "America's Best Hospitals". http://www.rti.org/pubs/abchmethod_2009.pdf. Accessed 30 July 2013

RTI International (2010) U.S. News & World Report Best Children's Hospitals 2010 Methodology. http://www.rti.org/pubs/abchmethod_2010.pdf. Accessed 30 July 2013

Sauer K, Kemper C, Kaboth K, Glaeske G (2010) BARMER GEK Heil- und Hilfsmittel-Report 2010. Asgard, St. Augustin

Schäfer T, Nolde-Gallasch A (1999) Modellversuch zur Beitragsrückzahlung bei den AOKs Lindau und Ostholstein – Bericht der wissenschaftlichen Begleitung. Technical report, University of Applied Science Gelsenkirchen, Bocholt (forthcoming, Federal Association of the AOK, Bonn)

Schäfer T, Neusser S, Lorenz C, Dörning H, Bitzer EM (2007) Krankenhaus-Rangfolgen nach Ergebnisqualität in der Hüftendoprothetik – Routinedaten mit oder ohne ergänzende Patienten-befragungen? – Teil 1: Routinedaten. GMS Med Inform Biom Epidemiol 3(1):Doc08

Schäfer T, Schneider A, Mieth I (2011) BARMER GEK Zahnreport 2011. Asgard, St. Augustin

Schneeweiss S, Schöffski O, Selke GW (1998) What is Germany's experience on reference based drug pricing and the etiology of adverse health outcomes or substitutions? Health Policy 44:253–260

Schneeweiss S, Seeger J, Maclure M, Wang P, Avorn, J, Glynn RJ (2001) Performance of comorbidity scores to control for confounding in epidemiologic studies using claims data. Am J Epidemiol 154:854–864

Schneeweiss S, Walker AM, Glynn RJ, Maclure M, Dormuth C, Soumerai SB (2002) Outcomes of reference pricing for angiotensin-converting-enzyme inhibitions. N Engl J Med 346:822–829

Schwabe U, Paffrath D (eds) (2010) Arzneiverordungsreport 2010. Springer, Berlin/Heidelberg

Schwartz FW, Busse R (2003) Denken in Zusammenhängen: Gesundheitssystemforschung. In: Schwartz FW, Badura B, Busse R, Leidl R, Raspe H, Siegrist J, Walter U (eds) Public Health. Gesundheit und Gesundheitswesen. Urban & Fischer, München/Jena, pp 518–545

Schwartz FW, Schach E (1989) Summary. In: Zentralinstitut für die kassenärztliche Versorgung in der Bundesrepublik Deutschland (ed): Die EvaS-Studie. Eine Erhebung über die ambulante medizinische Versorgung in der Bundesrepublik Deutschland. Deutscher Ärzte Verlag, Köln, pp 31–42

Schwarze E-W, Pawlitschko J (2003) Autopsie in Deutschland: Derzeitiger Stand, Gründe für den Rückgang der Obduktionszahlen und deren Folgen. Dtsch Arztebl 100:A2802–2808

Scott I, Campbell D (2002) Health services research: what is it and what does it offer? Intern Med J 32:91–99

Selby JV (1994) Case-control evaluations of treatment and program efficiency. Epidemiol Rev 16:90–101

Selmer RM, Kristiansen IS, Haglerød A, Graff-Iversen S, Larsen HK, Meyer HE, Bønaa KH, Thelle DS (2000) Cost and health consequences of reducing the population intake of salt. J Epidemiol Community Health 54:697–702

SF-36 Health Survey Scoring Demonstration. http://www.sf-36.org/demos/SF-36.html. Accessed 29 Mar 2011

SF-36 Psychometric Considerations. http://www.sf-36.org/tools/sf36.shtml. Accessed 29 Mar 2011

Shen Y, Ellis RP (2002) How profitable is risk selection? A comparison of four risk adjustment models. Health Econ 11:165–174

Shojania KG, Showstack J, Wachter RM (2001) Assessing hospital quality: a review for clinicians. Eff Clin Practice 4:82–90

Sonnenberg A, Delco F (2002) Cost-effectiveness of a single colonoscopy in screening for colorectal cancer. Arch Intern Med 162:163–168

Sörensen HT (2001) Routine registries. In: Olsen J, Saracci R, Trichopoulos D (eds) Teaching epidemiology. Oxford University Press, Oxford/New York, pp 99–106

Soumerai SB, Avorn J, Ross-Degnan D, Gortmaker S (1987) Payment restrictions for prescription drugs under Medicaid. N Engl J Med 317:550–556

Soumerai SB, Ross-Degnan D, Avorn J, McLaughlin JT, Choodnovskiy I (1991) Effects of Medicaid drug-payment limits on admission to hospitals and nursing homes. N Engl J Med 325:1072–1077

Spilker B (ed) (1995) Quality of life and pharmacoeconomic clinical trials. Lippincott-Raven, Philadelphia, PA

Statistisches Bundesamt (ed) (2000a) Gesundheitsbericht für Deutschland. Metzler-Poeschel, Stuttgart

Statistisches Bundesamt (ed) (2000b) Konzept einer Ausgaben- und Finanzierungsrechnung für die Gesundheitsberichterstattung des Bundes. Metzler-Poeschel, Stuttgart

Strauss DJ, Shavelle RM (2000) University of California life expectancy project. http://www.lifeexpectancy.com/index.shtml. Accessed 29 Mar 2011

Stuart A (1968) Basic ideas of scientific sampling. Charles Griffin & Co, London

Sudgen R, Williams A (1990) The principles of practical cost-benefit analysis. Oxford University Press, Oxford

Sukhatme PV, Sukhatme BV (1970) Sampling theory of surveys with applications. Asia Publishing House, London

Sullivan DF (1971) A single index of morbidity and mortality. HSMHA Health Rep 86:347–355

Sutton JH (2001) Physician data profiling proliferates. Bull Am Coll Surg 86:20–24

Tenney JB, White KL, Williamson JW (1974) National Ambulatory Medical Care Survey: background and methodology. In: National Center for Health Statistics (ed) Vital and health statistics series 2: data evaluation and methods research No. 61. U.S. Government Printing Office, Washington, DC

Thiemann DR, Coresh J, Oetgen WJ, Powe NR (1999) The association between hospital volume and survival after acute myocardial infarction in elderly patients. N Engl J Med 340:1640–1648

Torrance GW (1976) Social preference for health status. SocioEcon Plan Sci 10:129–136

Torrance GW (1987) Utility approach to measuring health-related quality of life. J Chronic Dis 40:593–600

Torrance GW, Thomas WH, Sackett DL (1972) A utility maximisation model for evaluation of health care programs. Health Serv Res 7:118–133

Townsend (2001) NHS resource allocation review: targeting poor health (vol I). In: Welsh Assembly's National Steering Group on the Allocation of NHS Resources. National Assembly for Wales, Cardiff

Tseng H-M, Lu J-F R, Gandek B (2003) Cultural issues in using the SF-36 Health Survey in Asia: results from Taiwan. Health and quality of life outcomes 1:72. http://www.hqlo.com/content/1/1/72. Accessed 29 Mar 2011

Tucker MA, Weiner JP, Abrams C (2002) Health-based risk adjustment: application to premium development and profiling. In: Wrightson C (ed) Financial strategy for managed care organizations: rate setting, risk adjustment, and competitive advantage. Health Administration Press, Chicago, pp 165–225

Tudor-Hart J (1971) The inverse care law. Lancet 1:405–412

Tudor-Hart J (2000) Commentary: three decades of the inverse care law. BMJ 320:18–19

U.S. Census Bureau (2009) Income, poverty and health insurance coverage in the United States: 2008. http://www.census.gov/newsroom/releases/archives/income_wealth/cb09-141.html. Accessed 29 Mar 2011

Van de Ven WPMM, Ellis RP (2000) Risk adjustment in competitive health plan markets. In: Culyer AJ, Newhouse JP (eds) Handbook of health economics, vol 1A. Elsevier/North Holland, New York, pp 755–845

van Mosseveld CJPM (2003) International comparison of health care expenditure. Statistics Netherlands, Voorburg/Herlen

Voß W (ed) (2003) Taschenbuch der Statistik, 2nd edn. Carl Hanser, München/Wien

Ware JE, Sherbourne CD (1992) The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 30:473–483

Weinermann JE (1971) Research on comparative health services systems. Med Care 9:272–290

Weinstein MC, Stason WB (1977) Foundations of cost-effectiveness analysis for health and medical practices. New Engl J Med 296:716–721

Wennberg JE, Freeman JL, Culp WJ (1987) Are hospital services rationed in New Haven or over-utilised in Boston? Lancet 1:1185–1189

Wennberg JE, Freeman JL, Shelton RM, Bubolz TA (1989) Hospital use and mortality among Medicare beneficiaries in Boston and New Haven. N Engl J Med 321:1168–1173

White T, Lavoie S, Nettleman, MD (1999) Potential cost savings attributable to influenza vaccination of school-aged children. Pediatrics 103:73e

Williams A (1974) "Need" as a demand concept (with special reference to health). In: Culyer A (ed) Economic policies and social goals. Martin Robertson, London

Williams A (2001) Science or marketing at WHO? A commentary on 'World Health 2000'. Health Econ 10:93–100

World Health Organization (WHO) (1947) The constitution of the World Health Organization. WHO Chron 1:6–24

World Health Organization (WHO) (2000) World health report 2000. Health systems: Improving performance. WHO, Geneva

Wright J (2001) Assessing health needs. In: Pencheon D, Guest C, Melzer D, Muir Gray JA (eds) Oxford handbook of public health practice. Oxford University Press, Oxford, pp 38–46

Wüthrich-Schneider E (2000) Patientenzufriedenheit – wie verstehen? Teil 1. Schweizerische Ärztezeitung/Bulletin des médecins suisses/Bolletino dei medici svizzeri 81(20):1046–1048

Zhao Y, Ellis RP, Ash AS, Calabrese D, Ayanian JZ, Slaughter JP, Weyuker L, Bowen B (2001) Measuring population health risks using inpatient diagnoses and outpatient pharmacy data. Health Serv Res 26(6 Part II):180–193

Ziese T, Neuhauser H, Kohler M, Rieck A, Borch S (2003) Flexible Ergänzung der Gesundheitssurveillance in Deutschland: Gesundheitssurveys per Telefon. Gesundheitsberichterstattung des Bundes, Berlin. http://www.sgw.hs-magdeburg.de/kurmat/goepel/hoge/ggf/grundlagen/yhtml/pdf/tel-survey.pdf. Accessed 29 Mar 2011

Ethical Aspects of Epidemiological Research 25

Hubert G. Leufkens and Johannes J. M. van Delden

Contents

H.G. Leufkens (✉)
Utrecht Institute for Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht,
The Netherlands

J.J.M. van Delden
Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht,
The Netherlands

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_40,

25.1 Introduction

> I think you need to give conscious consent to having any data, any personal data used, whether you are identified or not. That's certainly a right. That's your information, it's your medical history. Whether it's identified or not, you should control it.
> Patient 14, in Willison et al. (2003)

> Each public health policy, even a policy of inaction, implies some value or ethical judgment about the good of the individual and the community . . . .
> (Lee 2012)

These two quotes are about values and expectations, about perceived responsibilities, and about community benefits and individual rights in medical care and clinical or epidemiological research and reflect thereby compellingly the tensions, the paradoxes, the different views, and ethical aspects concerning biomedical research (Coughlin 2000). Epidemiology is part of the arena of biomedical research and is particularly focused on determinants of disease occurrences in populations. Ethics is the systematic analysis of values and norms (Weed and Coughlin 1999; Weed and McKeown 2001). Usually ethical reasoning and conduct are not issues that are at the top of an epidemiologist's menu chart (Beauchamp et al. 1991). In previous chapters of this handbook, we have seen that most epidemiological methods are non-interventional, e.g., observational by design, meaning that conventional ethical aspects of experiments with human beings (e.g., protocol review, randomization, placebos, informed consent) are not applicable as such. Many ethical committees have been struggling with the review of protocols of non-interventional studies because of the rationale and design of the study being directed at not influencing the "natural" disease course of patients, but at determining statistical inferences between various exposures (e.g., environment, drug treatment, and medical practice) and effects in the population in a non-experimental fashion. Observational epidemiology possesses the attractiveness, but also the practical paradox, of scientific investigation with a priori objective of not intervening in the normal course of the study object.

There have been several drivers within and outside the field of epidemiology that have changed the picture of ethical aspects significantly over the last decades. First of all, since the mid-1980s, the development and availability of automated record linkage databases, including biobanks, capturing both exposure and outcome data on an individual level, have raised questions about confidentiality of patient's medical records, authorizing access to person-specific information, and misuse of such databases (Knox 1992; Stjernschantz Forsberg et al. 2009). A second driver has been the debate about integrity and conflict of interests related to epidemiological research, in particular in cases of sponsored epidemiological studies and/or when the results of such studies were contradictory and subject of controversy and discourse. Finally, there is a growing interest of epidemiologists in biobanks where massive collections of biological materials and related data (e.g., tissue samples, cytological data, proteomic biomarkers) are assembled for diagnostic or therapeutic purposes and increasingly used for epidemiological research. The booming development of biobanks fuels ethical questions and debate about the design, conduct, informed consent, and competing interests between the individual patient, research objectives, and society as a whole (Stjernschantz Forsberg et al. 2009; Giesbertz et al. 2012).

As a consequence, the last decades have shown that concern about the ethical aspects of their research activities has become engaging epidemiologists as much as others who deal with public health, clinical decision making, and prioritizing and policy making in health care. Ethics guidelines have been prepared and accepted by several epidemiological organizations (Bankowski et al. 1991; IEA International Epidemiological Association 2007) in response to a growing awareness among epidemiologists that ethical conduct is essential to epidemiology. Basic principles of integrity, honesty, truthfulness, fairness and equity, respect for people's autonomy, distributive justice, doing good, and not harming have been made explicit. Essentially, the appreciation of these values has their origins in the follow-up of the Nuremberg trials, the UN Declaration on Human Rights, the Declaration of Helsinki, and many later declarations, guidelines, and codes of conduct. Basically, these declarations and guidelines reflect a major shift in current society from less priority to collective interests and benefits towards the primacy and protection of the individual (World Medical Association 2000; Coughlin 2000).

25.2 Drivers of Awareness of Ethical Aspects in Epidemiology

25.2.1 Surge of Automated Databases

One of the visionary founders of medical registries has been William Farr who understood already in the nineteenth century clearly the importance and potential of keeping person-specific records on diagnoses, medical treatments, environmental factors, and disease course (Farr 1875). Later various approaches of building and linking datasets for evaluating medical treatments and other determinants of health have been developed. In the early 1960s and 1970s, consistent and protocol-based medical record keeping became also recognized as an essential tool for clinical practice, and worldwide famous centers of clinical and epidemiological excellence like the Mayo Clinic or the Oxford Radcliffe Infirmary earned their appreciation mainly because they were champions in collecting and managing clinically relevant person-specific information in an era when paper charts, pencils, and several primitive collecting and retrieving machineries where state-of-the-art technologies (Gostin 1997). The introduction of advanced computer and information technologies changed that picture dramatically in the mid- and late 1980s. Storing, assembling, and linking clinical information became "push button" actions, and fascinating avenues for epidemiological research capturing data of hundreds of thousands, even millions, of individuals became feasible (Quantin et al. 1998). However, the "push button" nature and the big numbers involved of these developments gave rise to various ethical questions, in the beginning still vaguely phrased but later in very pronounced way on the table.

The overwhelmingness of the potentials of new information technologies and the speed of the developments have driven the need for a comprehensive balance sheet of all the social, political, and ethical aspects involved. Moreover, the owners and stakeholders of these automated databases are usually not health-care providers or professionals but third-party payers, e.g., health insurers, health maintenance organizations, or governmental bodies. These organizations have mostly not their origins in the professional and ethical environment of Hippocratic medicine and have invested in these data systems with other purposes (e.g., reimbursement of health-care providers, cost containment, risk management) than supporting medical practice (Gostin 1997).

The surge of automated databases has been stirred by the progress in record linkage techniques. Record linkage is the process by which pairs of correctly matched records of person-specific information are brought together in such a fashion that they may be treated as a single record for one individual (Herings et al. 1992). Record linkage provides a powerful tool in epidemiology in order to stratify exposures according to patient outcomes, e.g., bringing data together on food intake and cancer events, or exposure to sleeping pills and hospitalizations for hip fracture. Record linkage has driven the expansion of automated databases, and from an ethical point of view, there have been at least two major concerns (Herings et al. 1992; Kelman et al. 2002). First of all, the operational process of linkage of individual data from a number of different sources using a unique person-specific ID (identification) requires patient identification. Researchers in epidemiology have developed for that reason probabilistic approaches of record linkage, using sets of in itself not unique identifiers. However, it is believed by some opponents that this approach of record linkage may also violate data confidentiality rules, i.e., each pseudonymized method needs to be validated, and then the use of person-specific data (e.g., name or other ID) is essential (Tondel and Axelson 1999). A second concern related to record linkage has always been the fear that (non-) medical data (e.g., insurance status, lifestyle, sexual behavior, socioeconomic position) are built-in epidemiological data frames enhancing the feasibility of making unintended and/or undesirable statistical inferences that could cause damage or distress to individuals (Kmietowicz 2001). As a consequence, in the advent of a surge in automated databases, many countries both in North America and Europe have taken comprehensive legal action to assure the protection of personal privacy (Vandenbroucke 1992; De Vet et al. 2003).

Today's balance sheet of the role of automated databases in epidemiological research looks very positive. Cancer epidemiology, cardiovascular epidemiology, and pharmacoepidemiology are branches in epidemiology where such databases are key resources. They all have in common complex multivariate and time-dependent exposure-disease occurrences. Confidentiality of person-specific information is one of the most imperative ethical aspects to consider. From that perspective it is impressive to witness how certain countries (e.g., Denmark, Sweden) or environments (e.g., some of the health plans in the USA) have accomplished health-care record systems where unique codes enable to bring person-specific data from different sources together for research in a legally and ethically sound way (Andersson et al. 2011; Bazelier et al. 2012). A key factor for success seems to be active patient, consumer, and citizen engagement in the planning, design, and governance of such systems.

25.2.2 Integrity and Conflict of Interest in Epidemiological Research

In the “ideal” world, basic values of integrity, objectivity, respect, and independence should be key to every field of science. Committed to the discovering of the truth, researchers design, conduct, and report on study results (Levinski 2002). This notion gives the impression of science being a logical and unbiased human activity. Current society has long relied on scientists’ professional commitment to truth and honesty.

However, disclosure of, for instance, a case of fraud by a Dutch neurologist participating in the “European stroke prevention study 2” (ESPS-2), a multicenter stroke study, scandalized both the medical research community and the public (Hoeksema et al. 2003). The neurologist had committed fraud, in the sense that he had used names and fingered data of existing patients without these patients actually being enrolled in the study. The University of Connecticut in the USA announced clear misconduct by a vaccine expert who had falsified preliminary data in two grant applications (Malakoff 2003). The university removed the expert as head of the research center, and a series of lawsuits between the university and the vaccine researchers took place. We notice here two obvious cases of serious misconduct in biomedical science, e.g., doctoring of data. Other examples of questionable and unethical scientific behavior include apparent study sponsor-induced bias, as well known from research sponsored by the tobacco industry into the association between smoking and lung cancer (Barnes and Bero 1998), and at the very end of the spectrum, fraud, and falsification of data. We will come back to industry sponsoring of epidemiological research into drug effects. But there are other more subtle constraints to scientific integrity.

In 2002 Levinski gave a very personal and historical account on how he started as a medical researcher and reflecting visibly on major ethical questions, e.g., the protection and reimbursement of human research subjects, informed consent, disclosure of financial interests, prestige of the academic institution, and personal career building. The account also shows that ethical weighing can vary strongly over time. What we believed as being ethically acceptable in the past might not be today or vice versa:

> In 1963, before the advent of institutional review boards (IRBs), I was a young academic physician studying the regulation of sodium excretion by the kidneys. I paid medical students approximately $50 to serve as subjects for experiments involving only saline infusions and the collection of blood and spontaneously voided urine samples. I do not remember exactly what I told the students about the risks of the experiments but am quite certain that I characterized them as nominal. In one subject, severe phlebitis developed at the site of an intravenous infusion and required extensive therapy. The research project was funded by the National Institutes of Health. I had no possibility of financial gain from it. My primary motive was academic – the desire to advance knowledge about an important physiological mechanism with a bearing on clinical conditions such as edema. A potent secondary motive was to advance my career by publishing the results of the research and maintaining grant support – academic currency that buys prestige and promotion.
> (Levinski 2002)

For epidemiology, conflicts of interest related to research sponsoring are a very contemporary and controversial issue. Thompson has defined conflicts of interests

> as a set of conditions in which professional judgment concerning a primary interest (such as a patient's welfare or the validity of research) tends to be unduly influenced by a secondary interest (such as financial gain).
> (Thompson 1993)

Financial interests related to the tobacco or the pharmaceutical industry have been subject of intense controversy since decades. These industries have been always active in engaging researchers (as well as public media) for promoting messages, sometimes contrary to the available epidemiological evidence on health risks related to smoking or the use of medicines. In the field of epidemiology of drug effects, two archetypal cases have paved the pathway of debate and controversy on the ethics of research sponsoring, conflict of interest, and scientific (mis)conduct:

Cardiovascular Risks of Calcium-Channel Blockers In 1995 Psaty et al. reported in the Journal of the American Medical Association (JAMA) about a population-based case-control study among hypertensive patients in order to assess the association between first myocardial infarction and the use of antihypertensive agents (i.e., beta-blockers, calcium-channel blockers, angiotensin-converting-enzyme inhibitors, diuretics). The main result of the study was that the use of short-acting calcium-channel blockers, especially in high doses, was associated with an increased risk of myocardial infarction (Psaty et al. 1995). An intense controversy on the scientific validity of the study, the consequences for treatment of patients with hypertension, and the financial implications for the companies marketing calcium-channel blockers followed in both the medical literature and the lay press. A surge of commentaries, reviews, and additional papers on the topic emerged in the literature. Stelfox et al. (1998) evaluated the obviously visible signatures of the debate in the medical literature and demonstrated a strong association between authors' opinions about the safety of calcium-channel blockers and their financial relationships with those industries having an apparent interest in the hypertension market. Supportive authors had more financial ties with manufacturers of calcium-channel antagonists, while critical authors were much less likely to be involved in industry sponsoring and other financial connections with manufacturers. Although the paper of Stelfox et al. could be criticized for methodological reasons (lack of adjustment for dynamics of actions-reactions of time in the aftermath of the Psaty et al. paper), the overall message remains valid: there is and has been an association between ties with sponsors, choice of study questions, and, possibly, study results.

Venous Thrombosis Risk of Oral Contraceptives In the same year as the calcium-channel blocker controversy emerged 1995, several case-control studies reported on a twofold increased risk of deep vein thrombosis and pulmonary embolism in females using the so-called third-generation oral contraceptives relative to second-generation oral contraceptives (Skegg 2001). These findings engendered a surge of further (for the most part case-control) studies primarily driven by questions on possible confounding by indication (e.g., health user effect meaning preferential prescribing of third-generation oral contraceptives to females with more

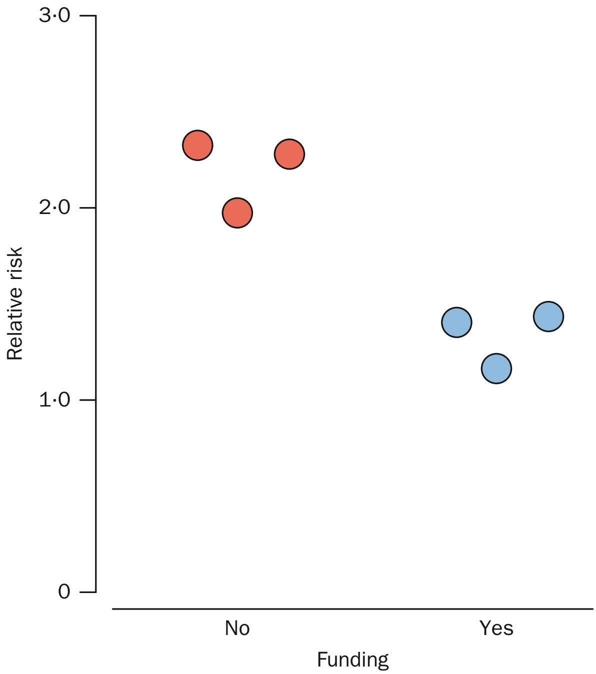


Fig. 25.1 Risk of venous thrombosis with third-generation contraceptives stratified for industry sponsoring (Source: Vandenbroucke 1998)

risk factors of cardiovascular disease) and biases related to the method of exposure ascertainment to oral contraceptives.

Many of these studies were sponsored by the pharmaceutical industry, and Vandenbroucke observed a contrast between the industry-sponsored studies reporting a relative risk of 1.5 or less and the non-funded studies consistently showing an increased risk of about 2.0 (Vandenbroucke 1998) (see also Fig. 25.1).

Answering the question whether this contrast is real, implicating that industry sponsorship is followed by biased research, is much more difficult to answer and is still subject of an ongoing debate. To illustrate the bewildering impression fuelled by the array of conflicting studies captured in Fig. 25.1, Vandenbroucke quotes a pharmacologist involved in early-phase studies for the industry:

> This might very well mean that industry-sponsored studies are the better ones. (Vandenbroucke 1998)

Like in case of the calcium-channel blockers controversy, a surge of commentaries and additional papers on the topic emerged in the literature and the public press. Moreover, as the topic was also subject for several court cases, the legal press covered the issue as well. This controversy has been one of the most striking examples in the last decade of how to find the truth in studying drug exposure-outcome associations, to unravel possible biases and confounding factors, dealing with study sponsor's interests, and at the same time to protect scientific integrity. Researchers are exposed to myriad pressures (e.g., balancing individual and institutional needs, search for professional recognition, and, sometimes, even rivalry). The science arena operates as a function of all the influences and pressures. Most progress to untangle the individual impact of all these factors has been made in demanding at least disclosure of all financial interests of the researcher by virtually all scientific journals and scientific communities (Levinski 2002). Epidemiologists need to continue to improve scientific and ethical conduct, to prevent unwanted conflicts of interest, and to be aware of the great financial interests of the parties involved (Beauchamp et al. 1991; Coughlin 2000). Various avenues to achieve this goal are either proposed or already in place: (1) codes of ethical conduct are adopted by virtually all professional and scientific societies, (2) the same holds for guidelines for disclosure of possible conflict of interests by authors submitting papers to medical-scientific journals, and (3) there is a surge both at the medical and other life science faculties, to include ethics classes in their standard curricula.

Full and transparent disclosure of any potential conflicts of interest affecting an epidemiological study to competent authorities, journal editors, or other regulatory bodies is part of virtually every code of guideline for conducting epidemiological research nowadays (von Elm et al. 2007; European Medicines Agency 2010). Each study should be evaluated in light of any declared conflicts and ensure that adequate means of mitigation are provided. When appropriate, it may also be required that a potentially conflicting interest be part of the information provided to the study subjects of the respondents. If a potentially serious conflict of interest cannot be adequately mitigated, the project should not be approved by the responsible committee.

25.2.3 Molecular Epidemiology and Genetics

Molecular epidemiology is a rapidly emerging field and in chapters ▶Molecular Epidemiology and ▶Statistical Methods in Genetic Epidemiology of this handbook, we have seen up-to-date accounts on scientific achievements and progress. A growing number of population-based molecular epidemiology studies have been set up to explore the roles of molecular factors (e.g., immune response profiling, blood clotting factors, enzymes), genomics (e.g., gene mutations as determinants of disease, gene-environment interactions), and other biomarkers (Maitland-van der Zee et al. 2000; Nuffield Council on Bioethics 2012; Kuehn 2011). Issues about participants' consent, confidentiality of information, and the feedback of findings have been widely addressed. Growing knowledge about molecular pathway-disease associations has led to new opportunities for testing, increasingly important as a guide to prevention, clinical management, and therapy. Tests are likely to vary in their predictive value, analytical and clinical validity, clinical utility, and social implications, e.g., access to and affordability of testing, insurance or employment discrimination, stigmatization, and long-term psychological harms from testing. Molecular epidemiology applying these tests is distinct from most other types of epidemiological research in that such biomarkers or genetic data obtained about an individual also may provide signatures of health about his or her relatives and person-specific future events (Khoury et al. 2011). For example, concerning the latter, the implications of a positive test for the breast cancer genes BRCA1 or BRCA2 mutation differ considerably for a woman who has not yet had children compared with one who has daughters who might be susceptible as well (Burke et al. 1997).

A pivotal and informative case in identifying and understanding the ethical aspects of these developments is the area of pharmacogenetics (Bolt et al. 2002). The increasing knowledge on the genome has resulted on unprecedented advances in understanding why individuals respond differently to drug therapy (Venter et al. 2001; Roses 2000). Pharmacogenetics focuses on the question of the extent to which genetic variants are responsible for inter-individual variability in drug response among recipients of a specific drug therapy. Few drug therapies are effective for everyone. The ultimate goal of pharmacogenetics is to shape therapy with available medicines in an individualized fashion, e.g., “tailor-made pharmacotherapy.” Pharmacogenetics integrates epidemiology, pharmacology, and genetics and is focused on an understanding of the genetic determinants of individual variability in drug therapy (Maitland-van der Zee et al. 2000). This research parallels the surge in discoveries of genes and protein expression patterns affecting the susceptibility to disease. There is ample evidence that certain disease susceptibility genes are also determining drug action and thereby therapy response.

The Nuffield Council on Bioethics (2012) has identified a number of ethical issues specifically raised by pharmacogenetics: (1) consent, privacy, and confidentiality; (2) management of information about response to therapy likelihood; and (3) implications of differentiating individuals into groups based on response to therapy likelihood. The key question in pharmacogenetics is unraveling the genetic traits of efficacy and/or safety of medicines. When that information is available, it can guide prescribers to select specific drugs or dosage schemes. It has been shown that on the one hand, male carriers of the apolipoprotein-E 44 variant are more prone to discontinue therapy with anticholesterol-lowering agents (Maitland-van der Zee et al. 2003). Although the precise mechanism underlying this association is still not known, prescribers, pharmacists, and patients can improve therapy knowing this risk factor of non-persistence by enhancing compliance with the regimen, tailor-made counseling, and the like. On the other hand, we know that apolipoprotein-E is also associated with various cardiovascular and neurological risks (e.g., Alzheimer disease). The level of evidence of the mentioned apolipoprotein-E associations is still subject of ongoing research, and all the three ethical issues mentioned by the Nuffield Council on Bioethics are visibly present in this case. This is particularly true in an area where we do not know today what kind of new genetic traits are discovered tomorrow and what kind of implications that has for already collected biological material (e.g., DNA samples). We see a surge in post hoc genotyping in both clinical and epidemiological research. This is feasible as individual genetics do not change over time, and when biological samples (blood, urine, or buccal cells) are still available, a major ethical question is whether the informed consent (maybe completed decades ago!) still holds for the current new situation. And what about the ethical questions provoked by genotyping cases and controls in, for instance, a case-control study revealing that certain study subjects carry serious susceptibility genes (e.g., BRCA1 or BRCA2 mutations)? Genotyping of the cases may be well covered by informed consent in the protocol, but this may be not valid for the controls sampled from the study base anonymously. And what about the "right not to know" of both the study subjects and their inherited relatives?

The application of genomic information to drug development also fuels ethical questions. Preferential inclusion of tested full responders into clinical trials increases the efficiency of such programs. However, such an approach would hide important information about the actions of the drug in other patients. In case the group of responders would be (too) small to develop the compound to an economically feasible medicine, the industry might decide to discontinue the project. On the one hand, the latter picture has led to the illustrious quote "Will all drugs become orphans?" (Maitland-van der Zee et al. 2000). On the other hand, drug development in many therapeutic areas (e.g., oncology, immune diseases) is increasingly driven by advances in cell biology and insight in molecular pathways of disease, making biomarker-enhanced targeting of drug treatment to the best responders, both in terms of efficacy and safety outcomes, more feasible and the most promising way forwards (Kuehn 2011).

25.3 Ethical Principles: Weighing Ethical "Benefits" and "Costs"

On the background of all the developments addressed so far, ethical principles are highly prevalent but in many cases badly defined, virtually invisible, or denied. Weighing of ethical "benefits" and "costs" is becoming an essential, additional perspective in designing and conducting sound epidemiological research (Nilstun and Westrin 1994). In the late 1980s, the Americans Beauchamp and Childress proposed four ethical principles in order to provide a more or less neutral, analytical framework to help doctors, researchers, and all others who are engaged in medical decision and policy making, when reflecting on moral issues that arise at work: respect for autonomy, beneficence, non-maleficence, and justice (Beauchamp and Childress 1989). Despite rapid and thought-provoking changes in medical technology and the practice of medicine, we believe that these four principles, plus attention to their scope of application, may encompass most of the moral issues that arise in today's health care and public health arena (Gillon 1994, 2003).

25.3.1 Autonomy

Autonomy is a widely discussed principle in bioethics and the word has several meanings. By and large, however, two focuses can be discerned: on the one hand autonomy can be perceived as a right to self-determination and on the other hand as an ideal of deliberated self rule. The first is about sovereignty and the second about authenticity. With respect to research ethics, autonomy is most visible in the practice of informed consent. Autonomy may be infringed if individuals are denied the right to choose whether or not to be enrolled in clinical or epidemiological research. Respecting people's autonomy requires consulting patients or other study subjects and obtaining their agreement before inclusion in a study. Medical confidentiality is an instrument to protect privacy, which in itself is based on the respect for a patient's autonomy.

Privacy refers to freedom of the person to choose for himself or herself the time and circumstances under which and, most importantly, the extent to which, his or her attitudes, beliefs, behavior, and opinions are to be shared with or withheld from others. Confidentiality refers to managing private information; when a subject shares private information with (confides in) an investigator, the investigator is expected to refrain from sharing this information with others without the subject's authorization or some other justification. Without confidentiality patients will be also far less open about all their personal concerns, symptoms, and other pieces of highly private information. Such information is very often critical to assign diagnoses and treatment scenarios to individual patients. This will have implications for clinical practice, but also for research. Study subjects should have more than enough reasons to trust researchers. Respecting autonomy also means not abusing this trust.

25.3.2 Beneficence and Non-Maleficence

The principle of beneficence means that health-care professionals and investigators have a responsibility to do good for those whom they treat. The traditional Hippocratic moral obligation of medicine is to provide net medical benefit to patients with minimal harm. Therefore, beneficence and non-maleficence are viewed as basic components of a balance sheet aiming at producing net benefit over harm. For epidemiology this means that a research project should add to the existing knowledge base on exposure-disease occurrences in order to treat populations of patients effectively and to prevent health hazards or even mortality in the community. In epidemiology the interests, and thereby the benefits, for the individual patient are less obvious, since often no treatment is offered. However, part of the benefit that communities, groups, and individuals may reasonably expect from participating in studies is that they will be told of findings that pertain to their health. Where findings could be applied in public health measures to improve community health, they should be communicated to the health authorities. In informing individuals of the findings and their pertinence to health, their level of literacy and comprehension must be considered. Research protocols should include provision for communicating such information to communities and individuals (Bankowski et al. 1991).

The principle of non-maleficence applied to epidemiology reflects the moral obligation not to do harm to study subjects. In many cases of research, it is still uncertain what the benefits are of a specific intervention (as this is part of the study question). The principle of non-maleficence teaches that at least participating in the study should do no harm and should involve only minimal risks. Likewise, epidemiological investigators studying activities that pose risks to the well-being of subjects are ethically obligated to propose to subjects with whom they interact any feasible steps that can be taken to minimize their exposure to risk. Furthermore, harm may occur, for instance, when scarce health personnel are diverted from their routine duties to serve the needs of a study, or when, unknown to a community, its health-care priorities are changed. It is wrong to regard members of communities as only impersonal material for study, even if they are not harmed. Ethical review must always assess the risk of subjects or groups suffering stigmatization, prejudice, loss of prestige or self-esteem, or economic loss as a result of taking part in a study.

25.3.3 Justice

This principle underpins the moral obligation of a fair distribution of burdens and benefits between people. One way of looking at justice is treating those with equal need equally. Justice can also be described as the requirement to act on the basis of fair settlement between competing claims and demands. Equity is at the heart of justice, and since centuries, people have argued about the morally relevant criteria for regarding and treating people as equals and those for regarding and treating them as unequals. This principle has become prominent in an era of cost-containment and rationing of health-care resources. Allocation of resources may conflict between several common moral concerns (e.g., individual access to and affordability of resources, fair distribution of scarcity, autonomy of professionals to make the best decisions for their patients). All concerns may be morally justified but not all can be fully met simultaneously. Epidemiology is the science of landscaping and explaining differences (in health, socioeconomic status, resources, risk factors) within populations and is thereby a critical "monitor" of (in)equity (Weed and McKeown 2001).

25.3.4 Balancing the Four Principles

Although, all the four principles together are seen as a comprehensive frame for moral reflection in medicine and epidemiology, we can observe a shift in emphasis and a greater prominence of autonomy as the leading manual for ethical conduct. This shift of putting the individual first is welcomed with mixed feelings and has made balancing individual rights with those of the whole society to become a key issue in contemporary Western society:

> Autonomy is, then, de facto given a place of honour because the trust of individualism, whether from the egalitarian left or the market oriented right, is to give people maximum liberty in devising their own lives and values.
> (Callahan 2003)

Table 25.1 Most important possible "benefits" and "costs" when the study is done

	Autonomy	Beneficence/non-maleficence	Justice
Study subjects	Costs	Benefit	Mixed
Physicians	Costs	Benefit	
Industry		Mixed	
Society at large		Benefit	Benefit

Nilstun and Westrin have proposed a model to cross the four ethical principles with the perspectives of each of the parties involved and then to assess and weigh the ethical "benefits" and "costs" for each individual party in the event the study is or is not conducted (Nilstun and Westrin 1994). Earlier in this chapter we have addressed the scientific and political debate about the risk of deep vein thrombosis and pulmonary embolism in females using the so-called third-generation oral contraceptives relative to second-generation agents. In 1999 Herings et al. published a follow-up study on this topic using anonymous exposure data related to females using one of these oral contraceptives and anonymous, but person-specific, outcome data on hospitalizations for either deep vein thrombosis or pulmonary embolism (Herings et al. 1999). The study confirmed the differential risk between the two categories of oral contraceptives and showed that the highest risk was in young females, newly starting with this contraceptive method. We use the study accessible through this chapter to illustrate and flesh out the model of Nilstun and Westrin.

The analysis starts with identifying the relevant parties (females using oral contraceptives, society at large, industry, prescribers). In Table 25.1 possible outcomes of an analysis of the most relevant "benefits" and "costs" are listed concerning the two dimensions of ethical principles and parties involved in the event that the study will be conducted. If the study is done, there are possible "benefits" for society at large, for prescribers, and for (other) women using oral contraceptives. For the industry the conduct of the study results in an ambiguous picture. Manufacturers of the second-generation oral contraceptives considered the study as "good news," for manufacturers of the third generation, the results of the study were less favorable. For industry as a whole, one may argue that every piece of science that contributes to the benefit-risk balance is advantageous, although this is not perceived like this in real life. Although this is reasonably understandable, it marks also the complexity and paradoxal nature of such multi-interest cases. Because the study confirmed earlier findings that most of the risk is concentrated in the very young users, this chapter provided important guidance to decision makers and young females in choosing the most suitable oral contraceptive.

With respect to the potential "costs," respect for autonomy (violating privacy, absence of individual informed consent) of the females and (possibly) the prescribers involved in the study is critical. Data used in the study were anonymous but person-specific, meaning that the investigator could not link the research data to any individual women. The linkage procedure of the Dutch Pharmaco-Morbidity (PHARMO) record linkage database has been internationally acknowledged (PHARMO data have been used in more than 200 studies) and brings community pharmacy and hospital data within established hospital catchments regions, together on the basis of patients' birth date, sex, and general practitioner (GP) code (yielding a sensitivity and specificity of linking person-specific data from two separate databases of over 98% each, which means that 98% are correctly linked) (Herings et al. 1992). In the same fashion, linkage between primary care data, population surveys, laboratory data, cancer, and accident registries has been achieved and applied in numerous epidemiological studies. But also, as previously stated in this chapter, record linkage using unique person-specific codes has shown to be very successful in building datasets for epidemiological studies (Andersson et al. 2011; Bazelier et al. 2012). A critical ethical issue has always been whether and how individual informed consent in such studies could be obtained. For sure, certain autonomy advocates argue that this should be accomplished. However, practical and methodological (those who refuse are mostly most relevant to the research) constraints would make individual informed consent virtually unfeasible. Instead, general informed consent in order to use the data for research purposes is very often obtained at the time a person enters the coverage area of a certain database. Coming back to the OC (oral contraceptives) and VTE (deep vein thrombosis and pulmonary embolism) study and looking at the principle of "beneficence," participating females have contributed (although not in conscious fashion) to the research and have taken their share in the solidarity of bringing together relevant data for solving an important public health problem.

Whatever the outcomes of such an exercise are, they provide a systematic frame for reflection and identification "where things can go wrong." The latter is a pivotal role of ethics in epidemiology (Coughlin 2000). Each preliminary idea of a study protocol should be accompanied with such an "ethical scan" not only for the purpose of moral justification of the research but also for reasons of improving the quality of the research. Experiences in coping with requirements to assure data privacy have been dominated mainly by technical (e.g., probabilistic linking, de-identification, introduction of random error on an individual level, but not on a population level) or procedural (e.g., standard operating procedures, good practice standards, security) dimensions (Roos and Nicol 1999). From a pragmatic view these dimensions may fully satisfy. However, ethical weighing of "benefits" and "costs" also includes critical reflection of the aims, deliverables, and consequences for the stakeholders (e.g., patients, physicians) involved. The latter goes beyond finding "smart tricks" to deal with privacy regulations or clinical trial directives.

As stated before, the ability to link person-specific clinical, exposure, and disease course data is a critical objective of epidemiology. Of all ethical issues and considerations, respecting autonomy by protecting privacy and confidentiality are the most crucial ones. Virtually all current legal systems in the Western world acknowledge the basic right of the patient to be assured that all his/her medical and personal data are confidential. Only in case of few well-defined exceptions, disclosure of person-specific information is allowed, e.g., prevention of serious risk to public health, order by a court of law in a crime case, and, under certain safeguards, scientific research. The tension between assuring personal privacy and access to medical data for epidemiological research has drawn ample attention from various stakeholders (individual patients, the public, politicians, health professionals, and the research community). In Table 25.1 possible violations of personal privacy related to either study subjects or physicians are classified as “costs.”

The scientific community of epidemiologists struggles with these two concepts and tries to convince politicians and policy makers of the importance of collective benefit to society from research with medical data and that we cannot rule out significant adverse effects to public health when epidemiological research has been made virtually impossible. Others take the pragmatic route using methodology that includes contemporary computer and statistical technology in order to build, within the framework of existing privacy legislation, aggregated, de-identified but person-specific, information. Court cases in several parts of Europe have concluded that the use of fully anonymous, de-identified patient data for the purpose of scientific epidemiological or clinical research is permissible under current law. In cases where it is not feasible to use primary data (collected directly from clinical practice for a specific, well-defined, purpose) in an anonymous fashion, informed consent should always be obtained. Epidemiological researchers may rely on access to non-anonymous medical records, but access to patient records for a research purpose requires individual patient informed consent.

The effect on research quality will be determined by the proportion of individuals who refuse consent, or in the case of large automated databases, who are simply not contactable. Researchers, cautioned by privacy advocates, very often overestimate participation proportions in consent procedures. There is growing evidence available that patients are willing to allow personal information to be used for research purposes. Several studies have shown that refusal to comply with consent procedures is most often not higher than in about one out of ten. A study from Canada suggests, however, that study subjects want to be actively consulted before the start of an epidemiological study where personal information is collected, whenever this is practically feasible (Willison et al. 2003). Secondary use of data (use of existing data for purposes other than those for which they were originally obtained) remains controversial as some interpreters of the law feel that secondary data use is prohibited because of the requirement for data to be used only for purposes compatible with those for which they were originally collected. In practice we see that this requirement is solved by obtaining general informed consent, although many researchers have sought exemption from the consent requirements in order to minimize selection bias, logistical obstacles, time consumption, and costs. Record linkage provides a powerful tool for the study of the natural history of diseases, the etiology of rare diseases, or the study of drug-effect associations with a (long) induction period between exposure and outcome (Herings et al. 1992; Sauver et al. 2011). The process of linkage of individual data from a number of sources, such as primary care records, secondary care records, prescribing and mortality data, requires patient identification and in many countries is not permitted because this is believed to breach data protection laws. In order to comply with confidentiality rules, researchers very often rely on medical staff providing them with a list of patients’ names and addresses to be used as a sampling frame. Although no medical details are given, the provision of names and addresses is clearly not anonymous, and this is likely to be in breach of most legal and ethical rules on the issue. Experiences so far in record linkage represent a patchwork of various approaches to link individual sets of drug or other exposure information and clinical data. We see, for instance, in Scandinavia, a surge in record linkage studies based on national unique identifiers (Andersson et al. 2011). In the absence of such linkage codes, researchers have to rely on other approaches for bringing separate datasets together to patient-specific linked records.

25.4 Ethical Issues Specific to Epidemiology

25.4.1 Unethical Quality of Research

This brings us to another ethical angle of epidemiological research, namely poorly conducted research. That kind of research will for sure not benefit patients or society, but may cause harm when it leads to unsubstantiated and wrong decision making in clinical practice or policy making in public health. Taubes (1995) has addressed this issue in his thoughtful paper on the limits of epidemiology where he accused the field for producing repeatedly exposure-outcome associations that do not hold very long because subsequent studies either contradict the findings or are unable to reproduce the main study results. Although scientific controversies are essential to progress and evolution in science, conflicting data and secondary turmoil in epidemiology do most often more harm than good (Vandenbroucke 1998; Skegg 2001). This means that "Good Epidemiology Practices" with the purpose to prevent or adjust a priori poorly designed studies are as important for quality assurance as for ethical reasons (IEA International Epidemiological Association 2007; European Medicines Agency 2010).

25.4.2 Global Bioethics and Inequity

A remaining, but not less important, ethical challenge for future epidemiological research is the gigantic inequity in global health. Large differences in disease burden, variable access to efficacious and safe medical technology, gaps in pharmaceuticals and health services are an enormous concern (World Medical Association 2000). Fighting against inequity in global health as a feature of modern medical and epidemiological ethics goes beyond the application of the Hippocratic oath. It is about prioritizing and about creating affordability and access, and epidemiology is and will be the pivotal science of fuelling policy making and strategic action with quantitative evidence for managing this global problem (Reich 2000; World Health Organization 2003). The triangle global inequity, epidemiology, and ethics contrast extremely with the ethics of individual "autonomy" of, for instance, a patient objecting against participating in a database study in the USA or Europe. So far, ethicists have been struggling with "prioritizing" ethical issues. Questions whether a disease burden of an orphan disease with a prevalence of less than

1:10,000 in the EU is, or should be, as equally important as the burden of a tropical disease affecting millions and millions of people are still difficult to address. There is increasing interest and debate in global bioethics in the context of how different countries deal with solidarity when resources for health care are scarce, when lack of equity means differential allocation of health services or interventions like pandemic influenza vaccination plans. DeBruin et al. (2012) have listed several factors contributing to access barriers including lack of accessible information, distrust of government and public health agencies and programming, lack of insurance and inadequate insurance coverage, poverty, transportation and mobility issues, and distance to care. Epidemiologists are often the providers of data on unequal access and the consequences of these for the individual and the population at large. We will face important challenges for epidemiologists and ethicists involved in these "contrasting" areas. Some criteria, however, have been developed already. Both the Declaration of Helsinki in its fifth amendment (2000) and the CIOMS guidelines for biomedical research (CIOMS 2008) state that research undertaken in populations with limited resources should be responsive to the health needs of the population. Moreover sponsors and investigators must ensure that products or knowledge generated by the research will be made reasonably available for the benefit of the population.

25.4.3 Epidemiological Determinism and Preventive Medicine

In the late 1990s, James Le Fanu, a UK-based general practitioner, wrote a reflecting and alluring book titled the "Rise and Fall of Modern Medicine" (Le Fanu 1999). In his book, the author is very critical about numerous features of contemporary medicine and health care, in particular about the role of epidemiology in medical education, knowledge building, and clinical practice.

Many of his arguments are close to the ethical questions arising from the determinism of epidemiology resulting in stratifying populations in categories of disease susceptibility, consequently leading to screening, "healthy" behavior, and preventive medicine. The promise of genomics and other molecular strategies for improving the practice of medicine must be pursued taking into account the fundamental ethical principles of autonomy, beneficence, non-maleficence, and justice. Because genetics are linked to family ties, the "right of not to know" for instance, goes beyond the individual judgment and decision making of the persons involved in the study themselves.

25.4.4 Precautionary Principle and Scientific Evidence

Closely related to the former issue of epidemiological determinism is the question how strong the evidence of the relationship between a hypothesized cause (i.e., environmental factor, medical intervention, drug treatment) and the effect should be before implementation and public health action is justified (Rogers 2003).

This is at the heart of the science of epidemiology, as we have seen in previous chapters, and there are many cases of “established” exposure-outcome associations which had to be revoked afterwards because new studies and data became available, e.g., reserpine and breast cancer, fenoterol and asthma death, or drinking coffee or tea and pancreas cancer (Fraser 1996; Spitzer et al. 1992; Lee et al. 2007). According to the precautionary principle, a principle widely embraced nowadays by politicians and consumer advocates, particularly in the area of assessing environmental risks, it is uncertainty that justifies and requires proactive measures and regulations and a reversal of the burden of proof. A manufacturer intending the marketing of a new product or a government planning to build a new power plant has the obligation to provide solid evidence on efficacy and safety of the innovation before authorization is granted. The pharmaceutical market knows this system already since the early 1960s, but also other areas of medical technology, food, and environmental exposures anticipate more proactive assessment of the benefits and risks. The application of the precautionary principle is controversial because, according to its opponents, it drives behavior of counterproductive risk-avoidance and defensive strategies in balancing risks and benefits of innovation. Assessment and prediction of health effects of any intervention depend on a synthesis of all available epidemiological and mechanistic evidences to produce a valid estimate of the likely effect. Epidemiology is an important scientific resource to fuel precautionary measures, but also to avoid that decision makers intervene in human behavior without sound evidence. From that perspective the adverse effects of the precautionary principle in terms of, for instance, exclusion of susceptible patients from certain medical technologies because proof of safety is still lacking (e.g., pregnant women, children), or neglecting and discontinuing research and development in specific risky areas, call for ethical reasoning.

25.4.5 Medical Ethics and Epidemiology

Singer and colleagues (2001) have identified a number of important drivers in medical ethics:

- New ethical challenges posed by advances in biotechnology
- Maturation of clinical ethics by strengthening the research base and developing graduate programs and fellowships
- Emphasizing the intersection between clinical ethics and health policy, including a focus on ethics of health-care institutions and health systems
- Increasing public education and involvement
- Developing the conceptual foundations of bioethics
- Changes in the doctor-patient relationship

Epidemiology is very close to clinical medicine, as epidemiologists provide scientific underpinnings of (1) the diagnosis, (2) etiology of the disease, and (3) the prognosis (and determinants of disease) in populations, both healthy and diseased. We anticipate that all major developments in medical genetics will have consequences for epidemiology ethics as well, directly or indirectly. But because epidemiology is frequently directed at the healthy part of the population, meaning those who are not (yet) ill, the field carries specific ethical responsibilities with respect to predictive competences, e.g., identifying risk factors and preventive medicine. Moreover, as stated before, there are not many scenarios in epidemiological research where study subjects individually can benefit directly from the study and/or the research results. Partly this is a consequence of the historical nature of, for instance, retrospective case-control or many cohort studies, the anonymity of the data, and the large numbers involved, making person-specific implementation of the study results to study subjects hardly feasible. These features contrast with clinical research with more options for direct patient benefit. Direct patient benefit (or harm) is also an important driver of the discussion on the ethics of placebo-controlled clinical trials in case there is an efficacious therapy available and not treating might harm the patient for sure, e.g., by severe worsening of the disease or mortality in oncology research, or suicide risk in evaluating antidepressive therapies (Storosum et al. 2001; Michels and Rothman 2003). The general rule is that placebo-controlled trials are only morally acceptable in the absence of proven effective therapy. However, this implies also that in certain areas where large placebo effects are expected, not having access to placebo-controlled data may result in weaker evidence to justify a sound benefit-risk decision.

25.5 Conclusions

Among many other factors, innovation in automated databases and biobanks, the surge in molecular and genetic knowledge, and controversies about scientific integrity, conflict of interest, and global bioethics have increased apprehension of the importance of ethical aspects in epidemiology. In the beginning, concern about loss of privacy has been a key driver of ethical questioning in epidemiology, and various techniques have been developed to cope with the confidentiality issue. The creation of unbiased person-specific histories (including both data on various exposure and outcomes) is a crucial requirement in epidemiology. Ethical weighing of “benefits” and “costs” can play an additional and relevant role as a vehicle for thoughtful reasoning (Beauchamp et al. 1991). Indeed, there have been expressed concerns about the various ways of misusing such data. In particular in the era of genetics and the increased interests of health insurers to reduce their business risks, there is a great need for prudence, protection, and careful weighing (Bolt et al. 2002; Nuffield Council on Bioethics 2012). Discovery of genes determining the response to various exposures including food, environmental factors, and medicines is an emerging area of genomic research as well and will produce new and intriguing ethical questions.

When considering the four ethical principles of beneficence, non-maleficence, autonomy, and justice on a scale of individual versus society as a whole, it is apparent that most of the “benefits” of epidemiological research can be attributed to the collective level (community, society) and that most of the “costs” fall down on the individual level. That makes epidemiology vulnerable for controversies where the individual-collective dimension is sensitive. The two examples of calcium-channel blockers and the users of oral contraceptives exemplified that participating study subjects themselves are virtually not benefiting from the study results. Current or future users of both drug categories however are in a much better position after the research has been done than before.

In this chapter we have used a number of pharmacoepidemiology case examples to illustrate and flesh out main ethical principles and questions relevant for epidemiology. The points addressed however have broader implications also for other fields in epidemiology. Pharmacoepidemiology as a discipline provides a challenging area of angles and perspectives to highlight certain ethical issues. Apart from the science in itself, the field is also heavily aligned with the economic, social, and political dimensions of developing and using medicines in a global scenery.

It is not a rare occurrence that epidemiological researchers perceive ethics as cumbersome, conservative, and antiscientific. Although these feelings may be justifiable in some cases, the ultimate balance sheet of more ethical weighing and reasoning will be positive. Ethical reasoning helps also to be concise in defining the research question, the design, and conduct of the study. Ethics and the linked formal and legal frameworks (e.g., scientific conduct guidelines, privacy protocols, ethical review board) undoubtedly have delivered in terms of quality push, critical reflection, and scientific enlightenment and will continue to do so in the future. The research community, clinical medicine, and patients all are major stakeholders in searching for and achieving mutual benefit from integrating ethics into epidemiological science (Gillon 2003).

References

Andersson D, Magnusson H, Carstensen J, Borgquist L (2011) Co-morbidity and health care utilisation five years prior to diagnosis for depression. A register-based study in a Swedish population. BMC Public Health 11:552

Bankowski Z, Bryant JH, Last JM (eds) (1991) Ethics and epidemiology: international guidelines. CIOMS, Geneva

Barnes DA, Bero LA (1998) Why review articles on the health effects of passive smoking reach different conclusions. JAMA 279:1566–1570

Bazelier MT, Bentzen J, Vestergaard P, Stenager E, Leufkens HG, van Staa TP, de Vries F (2012) The risk of fracture in incident multiple sclerosis patients: the Danish National Health Registers. Mult Scler 18(11):1609–1616

Beauchamp TL, Childress JF (1989) Principles of biomedical ethics, 3rd edn. Oxford University Press, New York

Beauchamp TL, Cook RL, Fayerweather WE, Raabe GK, Thar WE, Cowles SR, Spivey GH (1991) Ethical guidelines for epidemiologists. J Clin Epidemiol 44(Suppl 1):151S–169S

Bolt LLE, Delden JJM van, Kalis A, Derijks HJ, Leufkens HGM (2002) Tailor-made pharmacotherapy: future developments and ethical challenges in the field of pharmacogenomics. Center for Bioethics and Health Law (CBG), Utrecht

Burke W, Daly M, Garber J, Botkin J, Kahn MJ, Lynch P, McTiernan A, Offit K, Perlman J, Petersen G, Thomson E, Varricchio C (1997) Recommendations for follow-up care of individuals with an inherited predisposition to cancer. II. BRCA1 and BRCA2. Cancer Genetics Studies Consortium. JAMA 277:997–1003

Callahan D (2003) Principlism and communitarianism. J Med Ethics 29(5):287–291

CIOMS (2008) International ethical guidelines for epidemiological studies. Council for International Organizations of Medical Sciences (CIOMS) in collaboration with the World Health Organization (WHO), Geneva

Coughlin SS (2000) Ethics in epidemiology at the end of the 20th century: ethics, values, and mission statements. Epidemiol Rev 22:169–175

De Vet HCW, Dekker JM, van Veen EB, Olsen J (2003) Access to data from European registries for epidemiological research: results from a survey by the International Epidemiological Association European Federation. Int J Epidemiol 32:1114–1115

DeBruin D, Liaschenko J, Faith Marshall M (2012) Social justice in pandemic preparedness. Am J Public Health 102:586–591

European Medicines Agency (2010) ENCePP guide on methodological standards in pharmacoepidemiology. http://www.encepp.eu/public_consultation/documents/ENCePPGuideofMethStandardsinPE.pdf. Accessed 23 Aug 2012

Farr W (1875) Supplement to the 35th annual report of the registrar general. HMSO, London

Fraser HS (1996) Reserpine: a tragic victim of myths, marketing, and fashionable prescribing. Clin Pharmacol Ther 60:368–373

Giesbertz NA, Bredenoord AL, van Delden JJ (2012) Inclusion of residual tissue in biobanks: opt-in or opt-out? PLoS Biol 10(8):e1001373. Epub 7 Aug 2012

Gillon R (1994) Medical ethics: four principles plus attention to scope. Br Med J 309:184–188

Gillon R (2003) Ethics needs principles – four can encompass the rest – and respect for autonomy should be first among equals. J Med Ethics 29:307–312

Gostin L (1997) Health care information and the protection of personal privacy: ethical and legal considerations. Ann Intern Med 127:683–690

Herings RMC, Stricker BHCh, Nap G, Bakker A (1992) Pharmaco-morbidity linkage: a feasibility study comparing morbidity in two pharmacy-based exposure cohorts. J Epidemiol Community Health 46:136–140

Herings RM, Urquhart J, Leufkens HGM (1999) Venous thromboembolism among new users of different oral contraceptives. Lancet 354(9173):127–128

Hoeksema HL, Troost J, Grobbee DE, Wiersinga WM, van Wijmen FC, Klasen EC (2003) A case of fraud in a neurological pharmaceutical clinical trial. Ned Tijdschr Geneeskd 147(28):1372–1327

IEA International Epidemiological Association (2007) Good epidemiological practice (GEP) IEA guidelines for proper conduct in epidemiological research. http://www.ieaweb.org/. Accessed 11 May 2012

Kelman CW, Bass AJ, Holman CD (2002) Research use of linked health data – a best practice protocol. Aust NZ J Public Health 26:251–255

Khoury MJ, Gwinn M, Clyne M, Yu W (2011) Genetic epidemiology with a capital E, ten years after. Genet Epidemiol 35(8):845–852

Kmietowicz Z (2001) Registries will have to apply for right to collect patients' data without consent. Br Med J 322:1199

Knox EG (1992) Confidential medical records and epidemiologic research. Br Med J 304:727–728

Kuehn BM (2011) Scientists see promise and challenges in translating genomics to the clinic. JAMA 305(13):1285–1286

Le Fanu J (1999) The rise and fall of modern medicine. Little Brown, London

Lee LM (2012) Public health ethics theory: review and path to convergence. J Law Med Ethics 40(1):85–98

Lee JE, Hunter DJ, Spiegelman D, Adami HO, Bernstein L, van den Brandt PA, Buring JE, Cho E, English D, Folsom AR, Freudenheim JL, Gile GG, Giovannucci E, Horn-Ross PL, Leitzmann M, Marshall JR, Männistö S, McCullough ML, Miller AB, Parker AS, Pietinen P, Rodriguez C, Rohan TE, Schatzkin A, Schouten LJ, Willett WC, Wolk A, Zhang SM, Smith-Warner SA (2007) Intakes of coffee, tea, milk, soda and juice and renal cell cancer in a pooled analysis of 13 prospective studies. Int J Cancer 121(10):2246–2253

Levinski NG (2002) Nonfinancial conflicts of interest in research. N Engl J Med 347:759–761

Maitland-van der Zee AH, de Boer A, Leufkens HGM (2000) The interface between pharmacoepidemiology and pharmacogenetics. Eur J Pharmacol 410:121–130

Maitland-van der Zee AH, Stricker BH, Klungel OH, Mantel-Teeuwisse AK, Kastelein JJ, Hofman A, Leufkens HGM, van Duijn CM, de Boer A (2003) Adherence to and dosing of beta-hydroxy-beta-methylglutaryl coenzyme A reductase inhibitors in the general population differs according to apolipoprotein E-genotypes. Pharmacogenetics 13:219–223

Malakoff D (2003) The multiple repercussions of a fudged grant application. Science 300:40

Michels KB, Rothman KJ (2003) Update on unethical use of placebos in randomised trials. Bioethics 17(2):188–204

Nilstun T, Westrin CG (1994) Analysing ethics. Health Care Anal 2:43–46

Nuffield Council on Bioethics (2012) Pharmacogenetics: ethical issues. Report September 2003. http://www.nuffieldbioethics.org/pharmacogenetics. Accessed 23 Aug 2012

Psaty BM, Heckbert SR, Koepsell TD, Siscovick DS, Raghunathan TE, Weiss NS, Rosendaal FR, Lemaitre RN, Smith NL, Wahl PW, Wagner EH, Furberg CD (1995) The risk of myocardial infarction associated with antihypertensive drug therapies. JAMA 274:620–625

Quantin C, Bouzelat H, Allaert FA, Benhamiche AM, Faivre J, Dusserre L (1998) Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. Methods Inf Med 37:271–277

Reich MR (2000) The global drug gap. Science 287:1979–1981

Rogers MD (2003) Risk analysis under uncertainty, the precautionary principle, and the new EU chemicals strategy. Regul Toxicol Pharmacol 37(3):370–381

Roos LL, Nicol JP (1999) A research registry: uses, development, and accuracy. J Clin Epidemiol 52:39–47

Roses AD (2000) Pharmacogenetics and the practice of medicine. Nature 405:857–865

Singer PA, Pellegrino ED, Siegler M (2001) Clinical ethics revisited. BMC Med Ethics 2(1):1

Skegg DCG (2001) Evaluating the safety of medicines, with particular reference to contraception. Stat Med 20: 3557–3569

Spitzer WO, Suissa S, Ernst P, Horwitz RI, Habbick B, Cockcroft D, Boivin JF, McNutt M, Buist AS, Rebuck AS (1992) The use of beta-agonists and the risk of death and near death from asthma. N Engl J Med 326:501–506

St Sauver JL, Grossardt BR, Yawn BP, Melton LJ 3rd, Rocca WA (2011) Use of a medical records linkage system to enumerate a dynamic population over time: the Rochester epidemiology project. Am J Epidemiol 173(9):1059–1068

Stelfox HT, Chua G, O'Rourke K, Detsky AS (1998) Conflict of interest in the debate over calcium-channel antagonists. N Engl J Med 338:101–106

Stjernschantz Forsberg J, Hansson MG, Eriksson S (2009) Changing perspectives in biobank research: from individual rights to concerns about public health regarding the return of results. Eur J Hum Genet 17:1544–1549

Storosum JG, van Zwieten BJ, van den Brink W, Gersons BP, Broekmans AW (2001) Suicide risk in placebo-controlled studies of major depression. Am J Psychiatry 158(8):1271–1275

Taubes G (1995) Epidemiology faces its limits. Science 269(5221):164–169

Thompson DF (1993) Understanding financial conflicts of interest. N Engl J Med 329:573–576

Tondel M, Axelson O (1999) Concerns about privacy in research may be exaggerated. Br Med J 319:706–707

Vandenbroucke JP (1992) Privacy, confidentiality and epidemiology: the Dutch ordeal. Int J Epidemiol 21:825–826

Vandenbroucke JP (1998) Medical journals and the shaping of medical knowledge. Lancet 352:2001–2006

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001) The sequence of the human genome. Science 291:1304–1351. Erratum in: Science 2001 Jun 5; 292(5523):1838

von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative (2007) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 370:1453–1457

Weed DL, Coughlin SS (1999) New ethics guidelines for epidemiology: background and rationale. Ann Epidemiol 9(5):277–280

Weed DL, McKeown RE (2001) Ethics in epidemiology and public health, [I] technical terms, [II] applied terms. J Epidemiol Public Health 55:855–857 [I], 56:739–741 [II]

Willison DJ, Keshavjee K, Nair K, Goldsmith C, Holbrook AM (2003) Patients' consent preferences for research uses of information in electronic medical records: interview and survey data. Br Med J 326:373–378

World Health Organization (2003) World Health Report 2003. Shaping the future. World Health Organization, Geneva

World Medical Association (2000) The revised Declaration of Helsinki. Interpreting and implementing ethical principles in biomedical research. Edinburgh: 52nd WMA General Assembly. JAMA 284:3043–3045

Part III

Statistical Methods in Epidemiology

Statistical Inference

26

John F. Bithell

Contents

J.F. Bithell
St Peter's College, University of Oxford, Oxford, UK

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_54,

26.1 Introduction

This chapter provides an overview of the principal aspects of statistical inference as it is encountered in epidemiology. It makes minimal assumptions about the theoretical background and attempts to provide a simple introduction to both the frequentist and Bayesian approaches. The former still predominates in the applied literature, while theoreticians increasingly argue that it should be replaced by the latter, Bayesian approach. In this chapter, we take the view that these two approaches to inference should not be regarded as antagonistic, but rather as aiming at fundamentally different objectives, which may be regarded respectively as the analysis of data and the weighing of evidence – the latter typically derived from essentially different sources.

In Sect. 26.2, we discuss the essentially inductive nature of statistical inference and the important distinction between statements about the probability of a random event and statements about the uncertainty attaching to essentially unknown quantities such as model parameters and statistical hypotheses. In Sects. 26.3–26.5, we examine the classical modes of statistical inference: point estimation of parameters, hypothesis testing, and the construction of confidence intervals. The importance and role of likelihood is addressed in Sect. 26.6, while Sect. 26.7 gives a brief introduction to Bayesian methods of statistical inference.

26.2 The Nature of Statistical Inference

Statistical inference is the process by means of which we attempt to learn about the nature of random phenomena from observations on the real world. Although the machinery of the subject is mathematical, it is important to understand that the inferential process is inductive and not at all like the deductive processes of mathematics. Induction rather than deduction is essentially the means by which science progresses and by which we learn about the real world. In this sense, statistics is, as an academic discipline, much closer to the sciences than to mathematics.

The inferences of the physical sciences have a degree of logical certainty to them: if an observation is inconsistent with a hypothesis or model of the world, then we can at least deduce with certainty that our model is wrong in some respect, even though we cannot deduce that some other model is certainly correct. In problems requiring a statistical analysis, however, even the negative inference is less certain, since chance plays a role and our statements inevitably involve an element of uncertainty. The mathematically accepted way to quantify the uncertainty attaching to a statement is through the use of probability, conventionally measured on a scale of 0 to 1.

26.2.1 Probability and Uncertainty

In this chapter, we assume that the reader has a basic familiarity with the concepts of probability and probability distributions, which enable us to develop *models* for random phenomena; such models will usually have associated parameters[1] and a primary goal of statistical inference is to make inferences about the values of these parameters. Thus, for example, the proportion of patients in a particular group who are of a specified sex is purely descriptive, but to use this proportion as an estimate of an underlying probability enables us to make inferences about the magnitude of this probability and also to make predictions about future observations. To do this, we consider the observed proportion as a particular realization of a *random variable*, a mathematical entity whose probability distribution describes the behavior of the generality of similar observations.[2] Again, measuring the average blood level of some prognostic indicator in a sample of patients is purely descriptive, but when we use this average to estimate the mean of a probability distribution (which might be assumed to be normal, for example), we are estimating a parameter in a model which will enable us to make inferences about the population of patients as a whole and to make predictions about future observations. We will illustrate the ideas and methods in this chapter by reference to three running examples, which we now introduce.

The Binomial Distribution Our first example is the simplest model of all, that for dichotomous events. The binomial distribution gives us the probabilities of observing a given number x of specific outcomes or attributes (e.g., being male) out of a total of m independent dichotomous trials representing individual outcomes or attributes (e.g., the sex of patients in some group). It is assumed that this probability p is the same for each individual and that the determinations are independent of one another. It is then shown in elementary textbooks on probability theory that the required probability is given by the binomial distribution[3]:

$$\mathrm{P}[X = x] = \binom{m}{x} p^x (1 - p)^{m-x}\ ;\ x = 0, 1, 2, \ldots, m, \tag{26.1}$$

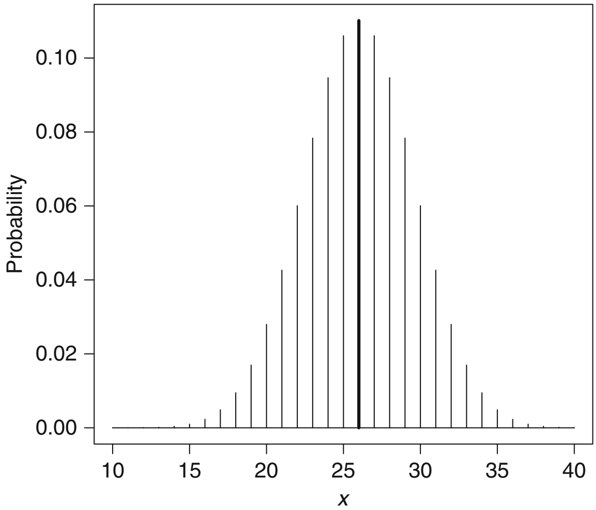
where $0 \le p \le 1$ and $\binom{m}{x} = \dfrac{m!}{x!(m - x)!}$ is the number of ways of choosing x items out of m without regard to order.

[1] This most useful word means a “quantity constant in (the) case considered, but varying in different cases” (Concise Oxford Dictionary). Its use to mean “boundary” (perhaps by confusion with “perimeter”), “limits” or “constraints” is to be deprecated.

[2] Mathematically speaking, a random variable is a function associating each possible outcome of an experiment with a particular numerical value, but at an elementary level, we can take it to be what it sounds like – a quantity that is variable from case to case in a manner that is intrinsically random.

[3] We follow the usual notational convention that a random variable is denoted by an uppercase letter and a specific value by a lowercase letter. P[·] denotes the probability of an event.

Fig. 26.1 Bar diagram showing the probabilities of the binomial distribution with $m = 52$ and $p = 0.5$



Example 26.1. Sex of meningitis patients
Bortolussi et al. (1978) report an investigation of newborn infants (under 31 days old) with meningitis in the Hospital for Sick Children in Toronto. Of the 52 infants they studied, 33 were male. It is reasonable to assume that each case presenting has the same probability p of being male and that the sexes of the children are determined independently, so the conditions for the binomial distribution apply. On the further assumption that boys and girls are equally frequent in the population and equally at risk of being affected by meningitis, we might suppose that $p = \frac{1}{2}$. Applying (26.1) with $m = 52$ and $p = 0.5$ gives the binomial distribution illustrated in Fig. 26.1. In particular, we can calculate that the probability of observing 33 males out of 52 children if $p = \frac{1}{2}$ is 0.017. The fact that this probability seems very small tells us little by itself about the plausibility of our model since it is clear from Fig. 26.1 that there are a lot of contributions to the distribution. On the other hand, it is also clear that this probability is a good deal smaller than the largest contribution, which is $\mathrm{P}[X = 26] = 0.110$.

The Normal Distribution In contrast to Example 26.1, we now consider the most important continuous distribution, conveniently represented by its density function, which may be depicted as a smooth curve with the mathematical formula:

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(x-\mu)^2/2\sigma^2) \,. \qquad (26.2)$$

The parameters μ and σ are the mean and standard deviation (SD) of the distribution, which measure respectively its center and its spread.

Example 26.2. Glycolated haemoglobin levels in a healthy population
Simon et al. (1989) reported a study of glycolated hemoglobin (HbA1c) levels in 3,240 healthy workers at France Telecom. Figure 26.2 illustrates the data as a histogram, together with the best fitting normal distribution, which has $\mu = 5.05$ and $\sigma = 0.54$. Judged visually, the shape of the empirical distribution is reasonably well approximated by the normal

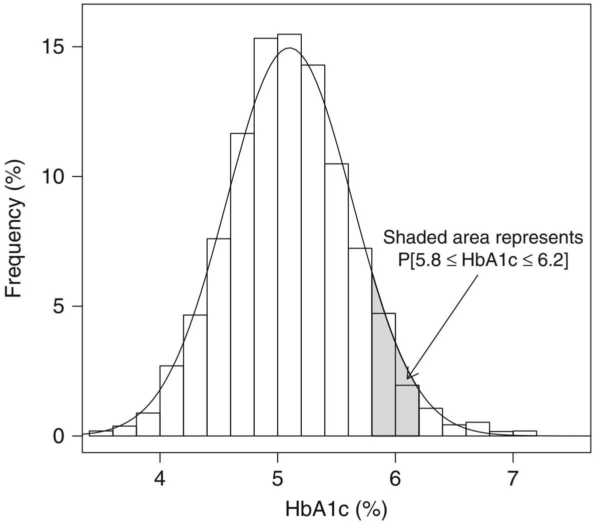


Fig. 26.2 Histogram of HbA1c data with best fitting normal density curve (Figure redrawn for this chapter with kind permission from Springer Science+Business Media from Diabetologia, see Simon et al. (1989), Fig. 1)

distribution. (In fact, a test of the kind introduced in Sect. 26.4.6 reveals a discrepancy resulting from the raised frequency in the right-hand tail, probably due to a small number of workers with poor glucose control.) The numerical value of the density function for a particular value of x is even less informative than in the discrete case and does not itself represent a probability, which necessarily lies between 0 and 1. For example, the most likely value for an observation is evidently at the mean of the distribution and here the density equals 1.37, which clearly could not be a probability. The function has to be interpreted as representing the probability of an observation falling between given values of x determined by the area under the curve between these values (see Fig. 26.2). The calculation of such probabilities is less straightforward than in the discrete case and proceeds by reference to the so-called *standard normal distribution* having $\mu = 0$ and $\sigma = 1$, for which the probabilities can be obtained from published tables or widely available computer programs.

The Poisson Distribution As a third example, we introduce the Poisson distribution, one of the most useful models of epidemiology. It describes the occurrence of events regarded as points in a continuum, for example, accidents regarded as point events in time or geographical space. In epidemiological practice, we do not often observe events this way; however, its use is much more widespread than this rather mathematical description might suggest because of a well-known result in probability theory stating that the Poisson distribution provides a very good approximation to the binomial distribution when the number of events m is large and the probability p of a positive outcome in each of the binomial trials is small. This means that the Poisson distribution provides such an approximation for the occurrences of all but the commonest diseases and for deaths over modest time intervals.

It may be objected that the individuals at risk of death or disease in a large population will generally have very different probabilities of succumbing. However, this is not a problem, since the Poisson distribution has the very important property that the sum of independent counts with different means will also have a Poisson distribution. If we imagine our population of individuals to be stratified, so that the individual probabilities of disease are small but very similar in each stratum, the Poisson distribution will provide an approximation to the number affected in each stratum. If we then sum these numbers, the result just stated ensures that the Poisson distribution will be a good approximation to the distribution of the total number of cases, even though the individual risks may show wide relative differences. Thus, all we need to do is to calculate or estimate the overall mean of the distribution, which is its only parameter. This extraordinary simplicity has far-reaching implications for statistical epidemiology and, indeed, for the whole of statistics.

If we let the mean of the distribution be denoted by λ, the probabilities are given by:

$$\mathrm{P}[X = x] = \mathrm{e}^{-\lambda}\lambda^{x}/x!,\ x = 0, 1, 2, \ldots\ ;\ \lambda > 0. \tag{26.3}$$

All the properties of the distribution are determined by the single parameter λ; in particular, its variance[4] is equal to the mean, λ.

Example 26.3. Emergency admissions to a London hospital
Figure 26.3 shows the distribution of the numbers of female emergency admissions on Saturdays and Sundays at a London teaching hospital in the year ending March 1966 (unpublished report). The mean number of admissions per day was 3.92, and the figure also shows the probabilities in the Poisson distribution with this mean. It will be seen that the agreement appears to be good; we shall examine below (Sect. 26.4.6) methods of assessing how good the fit is. In fact, the mean numbers for other weekdays differ, and we have achieved a good fit here by restricting observations to days with similar admission rates, namely days at the weekend. A more literal model for emergency admissions would consider a large number of individuals at risk of being admitted, though all with widely varying probabilities. But it is highly unlikely that we would have the required information to determine the different probabilities for these individuals and, as discussed above, the situation can be well approximated by the Poisson distribution even when the probabilities vary.

26.2.2 Modes of Inference

Examples 26.1 to 26.3 are typical of the use of probability models to describe data subject to random variation and to make predictions about future observations. All their statements are in the form of a conditional probability such as P[Observation(s) | Model assumptions], where the | indicates the operation of conditioning so that this expression reads “probability of specified observations given that the model assumptions are correct.” Statistical inference, however, requires an inversion of such expressions to read “plausibility of model assumptions given the observations made.”

[4] The variance is the square of the SD and is a measure of spread, or dispersion, that is more convenient than the SD for some purposes.

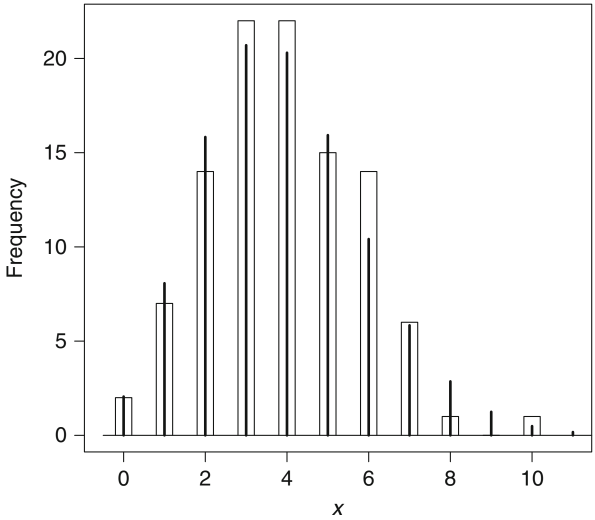


Fig. 26.3 Distribution of the daily numbers of emergency admissions (*unfilled blocks*) and the probabilities for the Poisson distribution with mean 3.92 (*thin black bars*)

The theory of probability is usually developed in the context of random events that are the outcomes of a sequence of experiments which could in principle be repeated indefinitely. Indeed, the easiest approach to probability is the *frequency approach*, which regards it as the limiting proportion of such outcomes in repetitions of a particular kind of experiment.

A moment's reflection, however, will convince us that a scientific statement is not like a random experiment: it is either true or false, and our ignorance of this does not endow it with the properties of a random event. Only by extending the notion of probability to apply to statements that are derived subjectively without reference to a potential frequency verification can we use it to quantify the uncertainty attaching to scientific statements. This is the extension that is espoused by the *Bayesian* school of statistical inference, a school that has gained ground appreciably over the last two decades (cf. chapter ▸Bayesian Methods in Epidemiology of this handbook). The kernel of this idea is Bayes' theorem[5] which provides a method of inverting probabilities by means of the formula:

$$\mathrm{P}[H|D] = \frac{\mathrm{P}[D|H] \times \mathrm{P}[H]}{\mathrm{P}[D]}. \tag{26.4}$$

As long as H and D stand for events in some random experiment, the formula is an uncontroversial consequence of the fundamental laws of probability. The issue becomes more difficult when D denotes some observations on the world, or Data, and H represents some explanatory hypothesis. For then the formula appears to

[5] This key result in statistical inference was first expounded by the Rev. Thomas Bayes (1701–1761), an English Presbyterian minister.

provide a probability for the hypothesis H, which we do not regard as random. It will be seen, however, that the right-hand side involves not only $\mathrm{P}[D|H]$, which is known as the *likelihood* of the data, but also $\mathrm{P}[H]$, that is, the probability of the hypothesis *before* we observe the data, known as the *prior probability* of H. Because it does not involve H, the denominator is of lesser importance, and we can summarize the law in a proportionality form:

$$\text{Posterior probability} \propto \text{Likelihood} \times \text{Prior probability}. \tag{26.5}$$

The more classical approach, often called *frequentist*, still prevails among most medical scientists, however, partly because of the feeling that introducing any subjective element into the argument weakens its scientific value, partly because the analyses are held to be easier to explain, though they are also easier to misinterpret. Because of their ubiquity, this chapter concentrates mostly on the frequentist modes of inference, and we deal in turn with the estimation of parameters, the testing of hypotheses, and the construction of confidence intervals. Each of these modes results in probability statements, but they are all of the form $\mathrm{P}[D|H]$, and their inferential value is correspondingly limited. There is also a tendency for such a statement to be misinterpreted as a statement reflecting $\mathrm{P}[H|D]$, often with very misleading results. The value of the frequentist approach lies in the appeal to the frequency interpretation of probability and its ability to show us the properties of a particular procedure, rather than its ability to assess the plausibility of a particular interpretation.

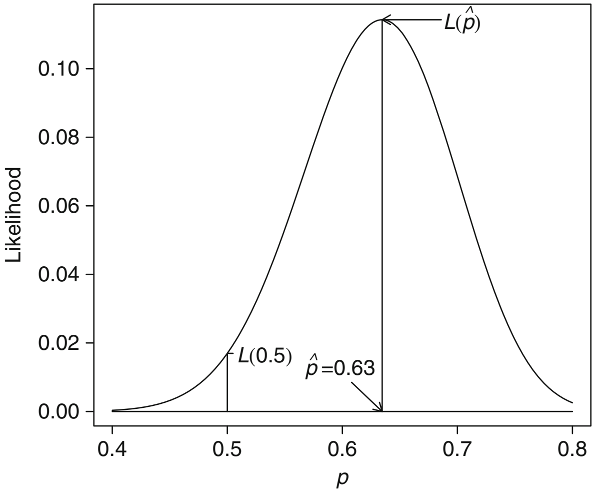
Although the frequentist approach appears to have the advantage of objectivity, it must be remembered that the result of a frequentist analysis expresses information only from the data being analyzed and this may be very modest compared with information we already have from other sources. As scientists, we might be content to weigh the evidence from our data with other evidence purely qualitatively, but this also introduces an element of subjective judgement, just as much as the Bayesian ambition to integrate evidence quantitatively.

Central to both Bayesian and frequentist methods is the concept of likelihood. Some statisticians even base their inferential ideas on it entirely. The concept is so important that we deal with it at the outset, before turning attention to the different modes of inference.

26.2.3 Likelihood and Plausibility

In fact, we have already introduced the likelihood in (26.4), and it will be seen that it appears there in the form of the probability of an observation given some model assumptions. The likelihood is indeed mathematically identical to the probability (mass or density) function, but expressed instead as a function of the unknown parameter with the data fixed, rather than as a function of the observation in a model with a given parameter value.

Fig. 26.4 Likelihood of the observation $x = 33$ from the binomial distribution with $m = 52$ as a function of the parameter p



Binomial Likelihood Thus, in the binomial case (Example 26.1), the likelihood for a given x is

$$L(p; x) = \binom{m}{x} p^x (1-p)^{m-x}\ ;\ 0 \le p \le 1. \tag{26.6}$$

The function is illustrated in Fig. 26.4 for the case $m = 52$, $x = 33$, and this should be compared with Fig. 26.1; it will be seen immediately that, like most likelihoods, it is a smooth function of the parameter of interest – in this case p. Instead of giving us the probabilities of different outcomes, different values now tell us the probabilities of the actual outcome for different values of the parameter. As with the probability function, the absolute magnitude of the likelihood tells us very little. But we have seen that the value of $L(0.5; 33) = 0.017$, which is among the smaller values available, and this suggests that if p really does equal one half, then something unusual has happened. Because unusual things happen rather rarely, it is entirely rational to find a larger value of p to be more plausible. Indeed, the most plausible value of p by this argument is at the maximum of the curve, which is easily seen to be at the value $p = 33/52 = 0.63$, for which $L(0.63; 33) = 0.114$. For obvious reasons, this value of p is termed the maximum likelihood estimate (MLE) of the parameter; it is usually denoted by $\hat{p}$. MLEs generally have good properties and play a major role in statistical inference.

Normal Likelihood Our second example follows the same lines. We first need to appreciate that the probability density of a sample of n independent observations $x_1, x_2, \ldots, x_n$ from the distribution is the product of expressions like that in (26.2); the likelihood is therefore the function of the two parameters μ and σ given by:

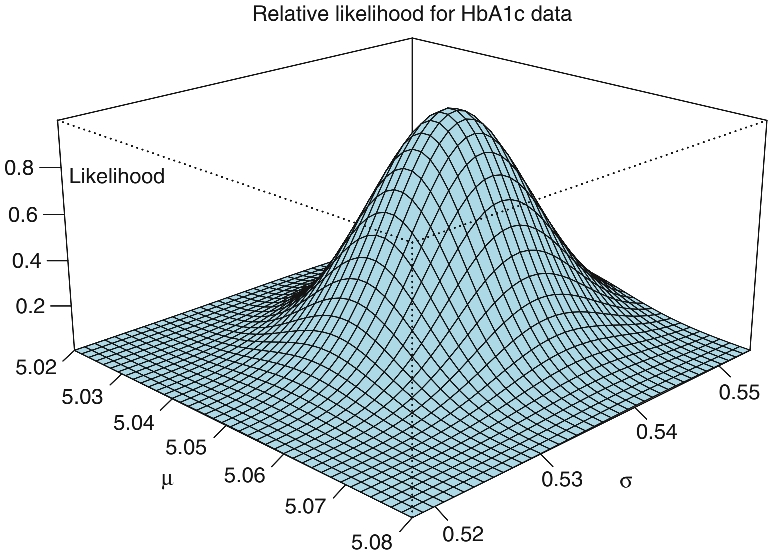


Fig. 26.5 Likelihood surface of μ and σ for HbA1c data having $n = 3,240$, $\bar{x} = 5.05$, and $\hat{\sigma} = 0.54$

$$L(\mu, \sigma; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-(x_i - \mu)^2/2\sigma^2\} . \tag{26.7}$$

It is more convenient to work with the logarithm of the likelihood, which of course has its maximum for the same parameter values as the likelihood itself. Thus, we write the log-likelihood as

$$\ell(\mu, \sigma) = k - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 = k - n \log \sigma - \frac{1}{2\sigma^2} \left(Q^2 + n(\bar{x} - \mu)^2\right), \tag{26.8}$$

where $Q^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$ and k is a constant not depending on the parameters, which is therefore irrelevant for the comparison of different values of ℓ.

This equation tells us important things about the likelihood in this case. In the first place, we need only know the values of $\bar{x}$ and Q^2 in order to compute it for any given parameter values. Hence, these quantities provide summaries of the data containing all the information about the parameters that is available in the likelihood. Such quantities are called *sufficient statistics*. Moreover, it is clear that, regardless of the value of σ, the function is maximized by choosing $\mu = \hat{\mu} = \bar{x}$, the sample mean. It is also quite straightforward to show that ℓ is maximized when $\sigma = \hat{\sigma} = Q/\sqrt{n}$, which is therefore the MLE of the SD of the distribution. Figure 26.5 shows the likelihood surface for the HbA1c data, for which $\hat{\mu} = \bar{x} = 5.05$ and $\hat{\sigma} = 0.54$.

As in the binomial case, relative values of L or different absolute values of ℓ indicate the relative plausibility of different parameter values. For example, with a postulated value of $\mu = 5.039$ (and maintaining the value $\sigma = 0.54$), we have a value of L half as large as the maximum value. Such comparisons give a measure of the plausibility of different parameter values, albeit not on a probability scale.

We can understand why the likelihood is so important if we reflect that it incorporates our understanding of the mechanism by which the data are generated; there is nothing in addition to this specification and the data themselves that gives any insight from our experimental observations. There are two provisos to this. The specification must be complete: sometimes it is hard to be sure which of two models with the same mathematical form for the likelihood actually did generate the data. It must also be correct since otherwise we may be seriously misled. Models are usually only approximately correct at best – for example, real data are never exactly normally distributed. Judgements about the seriousness of such departures from the ideal make up a large part of the business of statisticians and experience is often of more help than mathematical analysis.

26.3 Point Estimation

We have already met the MLE and the idea that the maximum likelihood approach can, in many situations, provide a good method of estimating parameters in a model. But we have not said what we mean by good; nor have we discussed comparisons with other methods. To do this within the frequentist framework, we need to make use of the concepts of an *estimator* and of its *sampling distribution*.

26.3.1 Sampling Distributions

An *estimator* is a function of our observations used for parameter estimation when we regard it as a random variable. This is in distinction to an *estimate*, which is an arithmetical quantity (or an algebraic formula for it) giving its value on the particular occasion we have observed. As a function of random variables, the estimator is also a random variable, and its long-run behavior can in principle be described by the use of probability distributions. Although it is not always easy to distinguish between an estimator and an estimate in a given context, it is nevertheless an important distinction, and it is good practice to observe it by using uppercase letters for the estimator and lowercase ones for the estimate. Thus, we write the estimator $\bar{X} = \sum_1^n X_i/n$ to represent the mean of a sample of observations when we wish to study how the sample mean behaves, but $\bar{x} = \sum_1^n x_i/n$ when we wish to describe how to calculate the sample mean from the observations $x_1, x_2, \ldots, x_n$.

In order to study the properties of an estimator, we must regard it as a random variable with a probability distribution determining the probabilities with which the estimate will have particular values. This probability distribution is known, quite generally, as the *sampling distribution* of the estimator and its properties tell us about the behavior an estimator would have if we could repeat our observations many times. In particular, it would seem desirable that the mean of the sampling distribution, around which successive estimate values would cluster, should be equal to the true parameter value. We say that an estimator with this property is *unbiased*. Unbiasedness sounds like a good property and so indeed it is, other things being equal.

However, we also need to consider the variability of our estimator about the mean of its distribution. One straightforward measure of this is the SD of the sampling distribution, and this is known as the *standard error* – often abbreviated s.e. – of the estimator. The smaller it is, the closer we expect our estimator to be to the unknown parameter – not with certainty on a given occasion, but on average in the rather precise sense of making the average squared error as small as possible. The standard error is an important concept, and estimated standard errors appear beside most published parameter estimates as adjuncts that enable us to judge their likely closeness to the unknown parameter values.

26.3.2 Estimating the Binomial Parameter

In the case of our binomial example, an obvious estimator of p is the sample proportion $\hat{p} = X/m$, and indeed we have already seen that this is the maximum likelihood estimator, for which we also use the abbreviation MLE, the distinction between estimate and estimator usually being clear from the context. Its sampling distribution is easily available since it is just a scaled version of the binomial distribution of X. It is an elementary result of probability theory that the mean of the binomial distribution of X is mp, from which we deduce that the mean of the distribution of $\hat{p}$ is indeed equal to p, that is, $\hat{p}$ is unbiased.

The theory of probability distributions also tells us that the SD of X is $\sqrt{mp(1-p)}$, so the standard error of $\hat{p}$ is the same quantity scaled by $1/m$, that is, $\sqrt{p(1-p)/m}$. We note that this is essentially unknown, since it depends on the parameter we are trying to estimate. Luckily, the function $p(1-p)$ changes rather slowly as p changes, so we may reasonably substitute our estimate into the formula to give the MLE with its s.e. as

$$\hat{p} = x/m \pm \sqrt{x(m-x)/m^3}. \tag{26.9}$$

This estimate of the s.e. will rarely be much in error unless we have small values of x or $m - x$. For the data in Example 26.1, we have $\hat{p} = 0.635 \pm 0.067$.

26.3.3 Estimating the Mean of the Normal Distribution

We have seen in Sect. 26.2.3 that the normal distribution has two parameters, μ and σ, with MLEs $\hat{\mu} = \bar{X}$ and $\hat{\sigma} = Q/\sqrt{n}$. It is mathematically more convenient for estimation purposes to regard the natural second parameter as σ^2, the variance of the distribution. MLEs are "invariant" under increasing transformations,[6] so it easily follows that the MLE of σ^2 is $\hat{\sigma}^2 = Q^2/n$. We will examine the properties of $\hat{\mu}$ and $\hat{\sigma}^2$ in turn.

In the first place, it can be shown that

$$\mathrm{E}[\bar{X}] = \mu \text{ ; and s.e.}[\bar{X}] = \sigma/\sqrt{n}, \tag{26.10}$$

where E[·] is the *expectation*, a term denoting the "mean of the probability distribution of." In fact, this important result applies with (almost) any *parent distribution* – that is, for any distribution from which the sample is drawn. From it, we can see that the inaccuracy in the estimate resulting from randomness decreases in proportion to $1/\sqrt{n}$. Thus, we need 100 times as many observations to improve our estimate by a factor of 10, a fact of life that is rather discouraging.

Another fact of life immediately clear from (26.10) is that the standard error of $\bar{X}$ depends on σ, which we will not usually know other than from the data and which therefore has to be estimated too. This is analogous to the situation with the standard error of $\hat{p}$, which requires a value of p. It is reasonable to use the MLE of σ for this purpose. The actual distribution of $\bar{X}$ is known in the case of a normal parent – it is in fact normal with the mean and SD given in (26.10).[7] Figure 26.6 shows the sampling distributions of $\bar{X}$ (a) for our actual data and (b) for a hypothetical sample of size 100; the first gives an idea of the precision[8] of our estimate obtained from the real data. The curves may also be compared with that for the parent (or population distribution) with $\sigma = 0.54$ (Fig. 26.2); this helps to emphasize the importance of the distinction between the SD of the parent and the standard error of the mean. To avoid confusion, it would be better when reporting the results of surveys to use the symbol ± to denote the s.e. rather than the SD, but the reader is warned that this convention is not always observed in the medical literature and there is a consequential risk of confusion.

[6] An increasing transformation of x returns a value $f(x)$ with the property that $f(x_2) > f(x_1)$ whenever $x_2 > x_1$.

[7] In other cases, it may be difficult or impossible to determine exactly, but the important Central Limit Theorem ensures that it will be very nearly normal in most circumstances.

[8] Technically, the precision is the reciprocal of the square of the s.e., that is, of the variance of the sampling distribution.

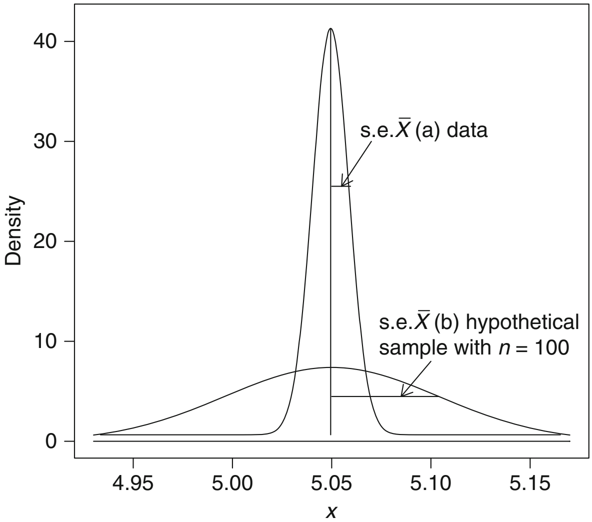


Fig. 26.6 Sampling distribution of $\bar{X}$ for samples of different sizes (**a**) for actual HBA1c data; (**b**) for a hypothetical sample of size 100

26.3.4 Estimating the Variance of the Normal Distribution

The same fundamental issues apply to the properties of $\hat{\sigma}$, but – as noted above – the theory works out better if we consider the properties of the variance estimator $\hat{\sigma}^2 = Q^2/n$. Elementary probability theory tells us that

$$\mathrm{E}[Q^2] = (n-1)\sigma^2, \tag{26.11}$$

so we can see that the MLE of σ^2 is biased, the unbiased estimator being $Q^2/(n-1) = S^2$, say. Unless n is small, the difference will not be important, but it is more usual to use the unbiased estimator than the MLE; we confine our attention to the properties of this estimator from here onwards.[9]

As with $\bar{X}$, the distribution of S^2 may be very hard to determine in general, though for a normal parent it has a distribution known as the chi-squared with $n-1$ degrees of freedom (d.f.).[10] For a general distribution, the standard error of S^2 is more difficult to determine than that of $\bar{X}$ since it depends on further properties of the parent distribution, but it is known for a normal parent and is given by s.e. $[S^2] = \sigma^2 \times \sqrt{2/(n-1)}$. We rarely make use of this, however, since the distribution of S^2 is skew – making its s.e. less informative – and in any case estimation of the variance of a sample is of secondary importance compared with estimating its mean.

[9] Interestingly, because the unbiased estimator has slightly greater variability, neither this nor the MLE has the smallest mean squared error – the estimator with this property is $Q^2/(n+1)$.

[10] The chi-squared distribution with k d.f., denoted by χ^2_k, may be defined as the distribution of the sum of squares of k independent variates having the standard normal distribution.

26.3.5 Estimating the Mean of the Poisson Distribution

The log-likelihood for a sample of n observations $x_1, x_2, \ldots, x_n$ from the Poisson distribution is given, from (26.3), by:

$$\ell(\lambda; x_1, x_2, \ldots, x_n) = k - n\lambda + \log \lambda \times \sum_{i=1}^{n} x_i, \quad \lambda > 0, \tag{26.12}$$

where, as before, k is a constant not dependent on the parameter. From this, we easily find that the MLE of λ is $\hat{\lambda} = \bar{X}$, the sample mean, and it can also be seen that this a sufficient statistic. The standard error of $\hat{\lambda}$ is $\sqrt{\lambda/n}$, and this is best estimated by $\sqrt{\hat{\lambda}/n}$, not by reference to the sample variance s^2. We can see too that the total count over the whole sample, that is, $T = \sum X_i$, is a sufficient statistic, and we know that it also has a Poisson distribution, with mean $n\lambda$. It follows that we could just as well estimate λ using the total count, the individual observations providing no further information as long as they really do have a Poisson distribution.

26.4 Hypothesis Testing

We turn now to a second major mode of frequentist inference, namely the testing of hypotheses. Typically, this is concerned with the testing of specific values of some parameters in a model like those of our first three examples. Very often, the values of these parameters are indicative of the data *not* establishing anything unusual, such as a hypothesis that there is no difference between two populations or that some parameter value has the value predicted by a simple model. For this reason the hypothesis tested is known as a *null hypothesis* and is universally referred to as H_0. A result leading to the rejection of H_0 is said to be *statistically significant* and the process of determining the test outcome is sometimes referred to as *significance testing*.

Some null hypotheses are so far reduced to fundamentals that they avoid specifying a parametric model at all. An example of this is the hypothesis that cases of a particular disease are no nearer to one another in space and time than would be expected by chance and it is possible to test this hypothesis by a procedure that makes no assumptions about the actual spatio-temporal distribution of the population at risk (cf. chapter ▶Geographical Epidemiology of this handbook). It is generally agreed that such tests are inherently less informative than tests in parametric models, but it can easily be realized that specifying a parametric model for the clustering of disease cases is very far from easy; such *non-parametric tests* are consequently attractive and very popular.

The actual procedure for testing a hypothesis appears in two related forms, which we deal with in turn.

26.4.1 The Formal Approach

The first approach follows that of Neyman and Pearson (1933), whose work in the second quarter of the twentieth century put the theory of testing on a firm foundation for the first time. Following the frequentist line of thought, they considered the repetition of a given testing procedure and defined the probabilities of making two kinds of error, the (type I) error of rejecting H_0 when it is true and the (type II) error of failing to reject H_0 when it is false. The probabilities of these two errors are usually denoted by α and β, and it is clear that if we choose a test procedure to reduce one of these probabilities, we will almost certainly increase the other so that there is a trade-off between the two types of error. To some extent, the choice of a test procedure then depends on the relative importance the experimenter attaches to each. We note, however, that the calculation of β is often far from straightforward since it requires the specification of an *alternative hypothesis* – usually denoted by H_1 – and it is generally unclear what this should be. We return to this point below.

This rather formal approach then proceeds along the following (rather simplified) lines:

1. A *simple* hypothesis H_0 is specified – that is, one that completely determines the distribution of the observed data, including the values of any parameters;
2. The set of all possible outcomes for the experiment (the so-called *sample space*) is identified and a *critical* (or *rejection*) *region* $\mathcal{R}$ is defined such that:
    a. The probability of the experimental outcome falling in $\mathcal{R}$ when H_0 is true equals a preassigned probability α, typically chosen to be 5%.
    b. Such an event casts doubt on H_0 in some scientifically significant way.
3. We reject H_0 if and only if this outcome is actually observed.

If we adhere to this procedure – which we assume to be specified in advance of observing the data – then we can be sure that, in the long run, we will make the type I error of wrongly rejecting H_0 on $100\alpha\%$ of occasions on which H_0 is true. It cannot be emphasized too strongly that this says nothing about how probable it is that we are wrong on a particular given occasion: all we can conclude from a rejection of H_0 is that something has happened that would be unusual if H_0 were true. The specification of $\mathcal{R}$ as a region suggests an almost unlimited number of different possibilities, but it frequently turns out that $\mathcal{R}$ can be defined in terms of a range of values for one particular statistic, such as a sample mean or a proportion.

26.4.2 Testing the Mean of a Normal Distribution

We have already studied the sampling distribution of the mean of a random sample from a normal distribution. Suppose we have such a sample and we wish to test the null hypothesis that the population from which the sample was drawn has a specified mean, μ_0, say. For simplicity, we will assume that the SD of the population

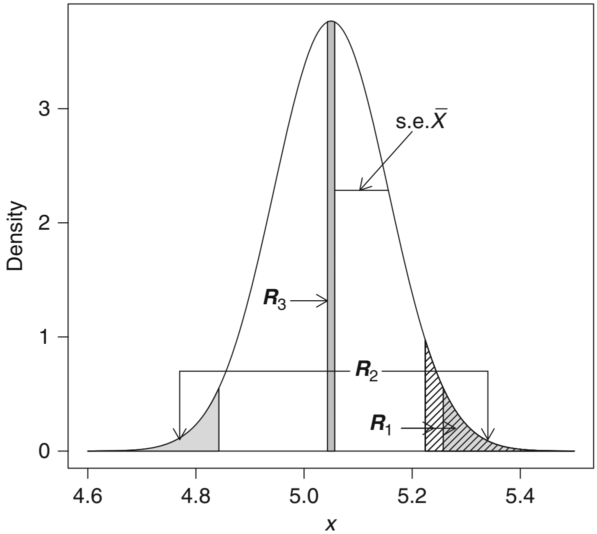


Fig. 26.7 Three different critical regions for a test of $H_0 : \mu = 5.05$ based on the HbA1c data (see text); the one-sided region (R_1) is *hatched*; the two-sided (R_2) and central (R_3) regions are *shaded*

distribution is known or specified and equal to σ_0, say. We know that the MLE of μ is $\bar{X}$, and we have also seen that it is sufficient, so it is a reasonable choice of test statistic, that is, we will use values of $\bar{x}$ to define a suitable critical region $\mathcal{R}$. Our critical region is then defined as a region on the $\bar{x}$ axis such that the probability that $\bar{X}$ falls in $\mathcal{R}$ is equal to our chosen α, which we will take to be 5%.

This does not in itself determine the region we should use; for example, Fig. 26.7 shows three critical regions of the right size, that is, three regions each having the property that the area under the normal density curve for $\bar{X}$ is $\alpha = 0.05$. There is no obvious choice for an alternative value of μ in this case, so we consider a *composite* alternative, that is, one that permits more than one possible value of μ. In particular, we might specify merely that $H_1 : \mu > \mu_0$. In this case, it would be sensible to choose the critical region $\mathcal{R}_1$ for $\bar{X}$ in the right-hand tail of the distribution under H_0, as shown hatched in Fig. 26.7. It is easily determined from tables of the normal distribution that, for $\alpha = 0.05$, the critical value of $\bar{x}$ is $\bar{x}_0 = \mu + 1.645\sigma_0/\sqrt{n}$; this region is termed one-sided.

However, we recall that this choice of test must be determined *before* we see the data, and it could have been the case that the mean for our patients would be *less* than μ_0. If we choose a one-sided test, such as that determined by $\mathcal{R}_1$, we must of course accept H_0 however small $\bar{x}$ may be. Such a result might be completely unexpected in the light of what we know about diabetes and blood glucose, for example, but it could nevertheless be very important to report it. Because nobody would expect an experimenter to ignore something important that was entirely unexpected, one would be right to be sceptical about the use of a one-sided test, and it is recommended practice in such situations always to use a test capable of detecting a deviation on either side of the null value.

Such a *two-sided* test is provided by the region $\mathcal{R}_2$, which has two parts (shown shaded). It is obvious that we now have to share the probability α between the two tails and – provided we keep α the same – this necessarily pushes the critical value(s) further out into the tails, thus reducing the chance that a positive deviation will fall in the critical region. The critical value of $\bar{x}$ is now 1.96 standard errors of $\bar{X}$ away from the mean.

The reader may wonder about the region $\mathcal{R}_3$ shown in Fig. 26.7. Such a region – although having the correct value of α – would be very unusual: clearly it will not detect values of μ very different from the hypothetical value. Indeed, it will reject H_0 if the sample mean is *too close* to μ_0; such a test might be appropriate if we suspected that the data were "too good to be true" in some sense; more usefully it could be an indicator of something wrong with the calculations in the test procedure, though it is unlikely that this alternative hypothesis would have been postulated a priori.

Example 26.4. Comparing HbA1c data with a standard
Czerniawska et al. (2008) studied the levels of HbA1c in 77 patients diagnosed as suffering from obstructive sleep apnea (OSA) and 22 controls. We will examine just the 26 OSA patients who were classified as not obese. To determine whether these patients have a "normal level" of glycolated hemoglobin, it might be thought reasonable to test the hypothesis that the patients have a distribution like that reported in Example 26.2, that is, a normal distribution with mean $\mu = 5.05$ and SD $\sigma = 0.54$. For this sample of 26 patients, a 5% two-sided test would have critical values of $\bar{x}$ satisfying $|\bar{x} - \mu_0| > 1.96\,\sigma_0/\sqrt{n}$, that is, outside the interval (4.84, 5.26). The non-obese OSA patients actually had a mean HbA1c of 5.4, which is therefore clearly in the critical region; we say that the OSA patients have a mean HbA1c significantly larger than the France Telecom population.

Testing the Difference of Two Means The same distribution theory leads to a test of the hypothesis that two samples come from normal distributions with the same mean, albeit that they may have different variances. It is an important property of the normal distribution that the sum or difference of two independent normal variates is also normally distributed, with mean and variance equal to the two respective sums. It follows that, if $n_1, n_2, \bar{x}_1, \bar{x}_2, s_1, s_2$ are the sizes, means, and standard deviations of samples from two normal distributions, then the statistic

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \tag{26.13}$$

has approximately a standard normal distribution, that is, with mean zero and variance one.

Example 26.5. Child IQ and maternal depression
Table 26.1 summarizes two samples of data on children's IQs according to whether their mothers had suffered from postnatal depression; the data were from Dr. Channi Kumar of the Institute of Psychiatry in London and are reported in full in Everitt (1994).

Table 26.1 Summary of IQs of children of depressed and non-depressed mothers

	No. of cases	Mean	SD	s.e.
Depressed	15	101.1	27.0	6.97
Not depressed	79	112.8	14.3	1.61

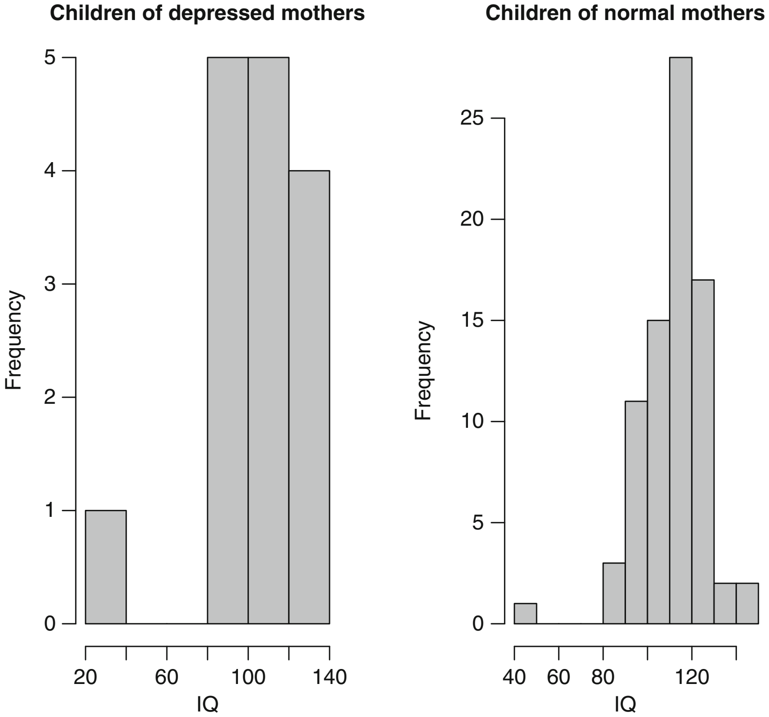


Fig. 26.8 Distributions of child's IQ in samples of depressed (*left*) and non-depressed (*right*) mothers

Applying (26.13) to these data gives $z = -1.64$, for the difference between the IQs of the children in the depressed and not depressed groups. We would not reject H_0 in a two-sided test, even when $\alpha = 0.10$. The distributions of the data are shown in Fig. 26.8, and we note that, although the samples are quite small, there is at least an appearance of non-normality in the data. We will consider alternative methods of analysis in Sects. 26.4.7 and 26.4.8, Examples 26.10 and 26.11.

26.4.3 The Power of a Test

The formal approach outlined above (Sect. 26.4.1) does not require us to specify the distribution of the observations when H_0 is not true. It is important to realize that whether we should really believe H_0 should depend on whether there are other hypotheses to be considered that are reasonably plausible in the light of the data. It may be, for example, that a small p-value can be matched by a comparably small value under a very plausible alternative "null" hypothesis, in which case it would seem more reasonable to ascribe the results to chance, at least in part. A non-significant result, on the other hand, might be due to a high value of β.

Calculating β in advance of an experiment is generally held to be an important part of the planning stage, and this is usually done in terms of the *power* of the test, defined as $1 - \beta$. A study with a low power is quite reasonably less likely to attract funding. Our problem is that, in most situations, we can only make an intelligent guess at an alternative hypothesis; the resulting power calculation will be hypothetical. Nevertheless, it is a useful discipline to make this calculation – perhaps for a range of different alternative hypotheses.

It is tempting to calculate the power after a study, using the parameter estimates as suitable values for the alternative hypothesis. This so-called "observed power" is sometimes used to assess the weight of the evidence for the null hypothesis, using the argument that a non-significant result might be due to low power, thus making it more likely that H_0 is true. This can be a very misleading argument and lead to what Hoenig and Heisey (2001) call the "power approach paradox." Imagine two experimental outcomes giving non-significant p-values, with $p_1 < p_2$. Because observed power is inversely related to the p-value, p_1 is indicative of higher "observed power"; this means that the p_1 experiment should have had a better chance of rejecting H_0 and the fact that it did not makes it more likely that H_0 is true, which contradicts the usual argument that a smaller p-value is associated with stronger evidence *against* H_0. In this rather unsettling example of flawed statistical reasoning, the essential error is the familiar confusion of the assessment of the probability characteristics of the testing procedure with the weighing of evidence from a particular analysis, though this manifestation of the confusion is heavily disguised and not generally recognized. The relative plausibility of different hypotheses is better quantified using the ideas discussed in Sect. 26.2.3 and again in Sects. 26.6 and 26.7.

In our description of the testing process, we appear to have conjured up a procedure largely on intuitive grounds. There is, however, an important result in mathematical statistics (the Neyman-Pearson lemma) which provides – at least in simple situations – a method of identifying the best $\mathcal{R}$ in the important sense of minimizing β for a fixed α. In particular, our choice of $\bar{X}$ and the critical region $\mathcal{R}_1$ is guaranteed to provide the most powerful test of H_0 against any one-sided alternative of the form $\mu = \mu_1 > \mu_0$. The situation is more complicated for other testing situations, but the lemma is nevertheless important, not least because it demonstrates the central role played by the likelihood in constructing a best test. In the normal case, the power for a given μ_1 is easily calculated; Fig. 26.9 illustrates how this can be done by considering the distribution of $\bar{X}$ under H_1 for a particular alternative value of μ_1. Thus, for example, we have a 65.6% chance of detecting an actual mean HbA1c of 5.3 in a two-sided test with $\alpha = 0.05$. If we repeat the calculation for many different values of μ_1, we generate the *power functions* shown in Fig. 26.10 for the one-sided and two-sided tests.

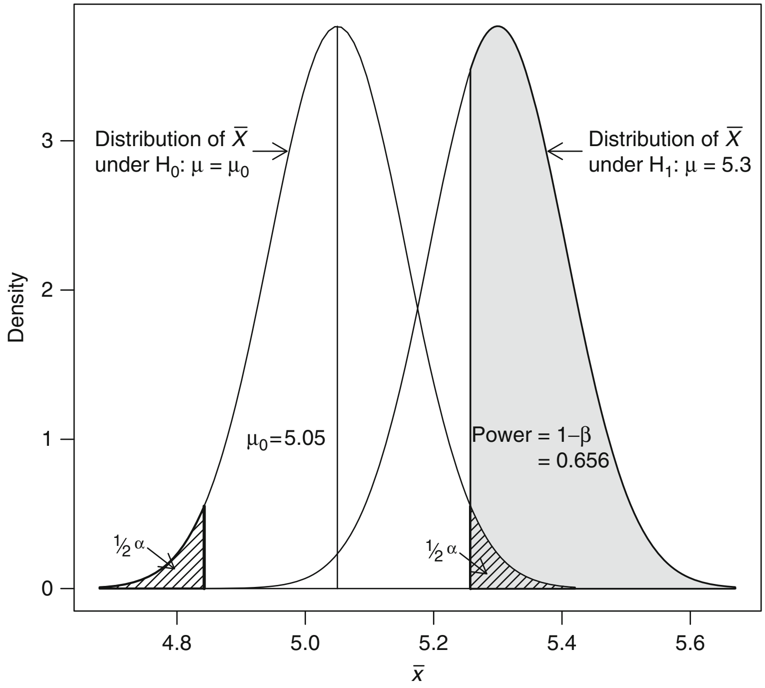


Fig. 26.9 Power of the two-sided 5% test of $H_0 : \mu = 5.05$ against $H_1 : \mu_1 = 5.3$ (HbA1c data)

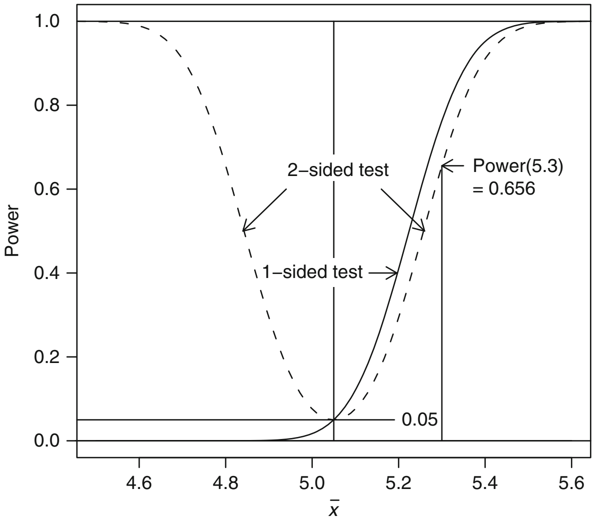


Fig. 26.10 Power functions for the one- and two-sided tests of the HbA1c data: test of H$_0$: μ = 5.05. The latter is indicated by the *broken line*

The normal model is extremely useful and is easily adapted to testing parameters in a wide variety of other models, using the fact that many parameter estimates are approximately normally distributed with standard errors that can be estimated reasonably well.

26.4.4 The p-Value Approach

One of the objections often raised against the formal Neyman-Pearson approach to testing hypotheses is that it is too "hard and fast." When we reject a null hypothesis, we are not establishing a logical statement since our conclusions inevitably involve uncertainty. It seems ridiculous to come to radically different conclusions according to whether our value of a sample mean is 1.95 or 1.97 standard errors bigger than μ_0. Arguably, we need a quantitative measure of the plausibility of H_0. Such a measure often advocated is known as the p-value. This is defined in the normal case as $\mathrm{P}[\bar{X} \geq \bar{x}_{\text{obs}}|H_0]$, that is, the probability that – if H_0 is true – we would observe a value of $\bar{X}$ at least as different from μ_0 as $\bar{x}_{\text{obs}}$, the value we actually did observe. In our Example 26.4, we have $\mathrm{P}[\bar{X} \geq 5.4] = 0.000475$; a two-sided test therefore has a p-value of 0.00095, and we immediately conclude from the fact that this is less than 5% that we would reject H_0 at the level $\alpha = 0.05$. We have the further information that we would have rejected H_0 even if we had specified a much smaller value of α, say 0.1%.

26.4.5 Discrete Distributions

When we have a discrete distribution for our test statistic, it will generally be impossible to find a critical region $\mathcal{R}$ with a probability α exactly equal to some prespecified value. In this situation, it becomes convenient to use the p-value approach since we can still calculate the probability in the tail of the distribution. The resulting p-value tells us whether our statistic would be in a critical region of a given size if it were possible to construct one. We illustrate these ideas by continuing our binomial Example 26.1.

Example 26.6. Testing the probability in the meningitis data
An obvious null hypothesis for the meningitis data is $H_0 : p = 0.5$, that is, the assumption that males and females are equally likely, which we made in Example 26.1. It is a familiar result that the binomial distribution is well approximated by the normal distribution with the same mean and variance. We can use this to obtain a good approximation to the p-value by treating the sample proportion X/m as a normal variate with $\mu = p = 0.5$, $\sigma = \sqrt{\frac{1}{2} \times \frac{1}{2} \times \frac{1}{m}} = 0.0693$. Hence, the probability of observing a value of $\hat{p} = X/m$ greater than or equal to the observed $33/52 = 0.635$ is approximately the same as the probability that a standard normal variate (i.e., one with zero mean and SD $= 1$) is greater than or equal to $(0.635 - 0.5)/0.0693 = 1.94$. The probability of this can be determined from tables to be 0.026. This should be doubled to give a two-sided test since we have no a

priori reason to suppose that males are more frequent in this context. We are thus led to the conclusion that we should not reject H_0 at the 5% level and we can quote a p-value of $P = 0.052$.

Alternatively, we can carry out an *exact test* by simply calculating $\mathrm{P}[X \geq 33|p = 0.5] = 0.0352$, giving a two-sided p-value of 7%. This exemplifies the fact that the normal approximation underestimates the exact p-value quite seriously for moderately small values of m, p, or $1 - p$. In fairness, the approximation would be considerably improved if we applied a *continuity correction*, which consists in reducing the absolute difference $|0.635 - 0.5|$ by $1/2m$; applying this adjustment gives a one-sided p-value of 0.0357, in good agreement with the exact calculation. The adjustment tends to overcorrect (and is therefore conservative) when m is small or p is near zero or one, but the relative error is substantially smaller than for the uncorrected approximation.

The Poisson Distribution Testing a simple null hypothesis specifying that one or more observations come from a Poisson distribution with a prescribed mean proceeds in a similar manner to the equivalent problem for the binomial distribution. Use of a normal approximation needs care when the information is limited. As a rule of thumb, the expected total of the observations should equal at least 100 to obtain reasonable accuracy in the tails; otherwise it is safer to conduct an "exact" test, summing the probabilities in the tails; see Armitage et al. (2002, Sect. 3.8) for comparisons between the exact and approximating values.

Although full information on a sample of n values $x_1, x_2, \ldots, x_n$ gives no more information about λ than the total $\sum x_i$, it does provide the possibility of checking the Poisson assumption by comparing the sample mean $\bar{x}$ and variance s^2. The ratio of these two statistics should be about one, and we can test it by forming $D = (n-1)s^2/\bar{x} = \sum(x_i - \bar{x})^2/\bar{x}$; this has an approximately chi-squared distribution with $n - 1$ degrees of freedom (d.f.) under the null hypothesis that the observations come from a Poisson distribution. We can use this to check our judgment that the emergency admissions data in Example 26.3, Fig. 26.3, are distributed in accordance with a Poisson distribution.

Example 26.7. Dispersion test for emergency admissions data

The sample of daily emergency admissions in Example 26.3 has mean 3.92 and variance 3.39. The "dispersion test" statistic described above is therefore $103 \times 3.39/3.92 = 89.1$. We can test it by reference to the chi-squared distribution with 103 d.f., which has an upper 5% tail point of 127.7, so that the mean and variance are not significantly different, testing at the 5% level. We justify using a one-sided test on the grounds that there is no plausible reason why the variance should be less than the mean. A variance greater than the mean, however, is very common, and it is indicative of the different observations having expectations that vary for one reason or another. In our case, for example, there could have been a seasonal effect leading to an inflation of the variance. Such data are termed *overdispersed* and the phenomenon is sometimes called *extra-Poisson variation*. It is not, however, necessary to abandon the Poisson distribution itself as an underlying mechanism; it will occur wherever occurrences of the phenomenon being counted are truly independent of each other. Infections or genetic diseases are likely to violate the independence requirement, but other diseases generally do not.

26.4.6 Goodness-of-Fit Tests

The dispersion test described above is an example of a goodness-of-fit test; such tests are important because can they provide a check on the assumptions of our model. There are many such tests and we illustrate here the elementary chi-squared test, attributed to Karl Pearson,[11] which is used for comparing a set of frequencies with expectations computed under a suitable model.

We suppose that we have a collection of k counts $f_1, f_2, \ldots, f_k$, as for example in a histogram. We compute a matching set of expectations $e_1, e_2, \ldots, e_k$ under assumptions about the origin of the frequencies that we wish to test. It will often be helpful to arrange the frequencies in the form of a table like Table 26.2; in more complicated situations, the same basic test can be applied to multi-dimensional tables. The statistic is calculated as

$$X_\nu^2 = \sum_{i=1}^{k} (f_i - e_i)^2 / e_i, \tag{26.14}$$

which has approximately a chi-squared distribution with ν d.f., where $\nu = k - c$ and c is the number of constraints between the frequencies and their expectations. There will usually be at least one such constraint since we will calculate the $\{e_i\}$ to ensure that $\sum e_i = \sum f_i$. Further constraints will arise in various ways, for example, from the estimation of parameters of a distribution determining the $\{e_i\}$. For the chi-squared approximation to work well, we should ensure that the expectations are not too small and this can be achieved by judicious grouping; a lower limit of 5 is generally regarded as being quite safe.

Example 26.8. Goodness of fit of emergency admissions data
Table 26.2 shows the frequencies displayed in Fig. 26.3 for Example 26.3, together with expectations calculated by multiplying the probabilities of the Poisson distribution by the total number of observations, 104. The tail of the original distribution was truncated as shown, and for testing purposes, the first two cells were also grouped in order to ensure that there are no expectations much smaller than five. The resulting Pearson's chi-squared statistic is 3.53, and this has $8 - 2 = 6$ d.f. since we have estimated the mean of the data. Since $P = 0.74$, this clearly confirms the impression that the fit of the data to the Poisson distribution is very good. Neither this test nor the dispersion test (Example 26.7)

Table 26.2 Daily female admissions to a London teaching hospital on Saturdays and Sundays and the frequencies expected in a Poisson distribution with the same mean

No. of admissions:	0	1	2	3	4	5	6	7	8+
Observed	2	7	14	22	22	15	14	6	2
Expected	2.06	8.07	15.83	20.70	20.30	15.93	10.42	5.84	4.86

[11] Karl Pearson (1857–1936) effectively founded the discipline of scientific statistics at University College, London, in the first quarter of the twentieth century; his son E.S. Pearson was the first Professor of Statistics in the same college.

rejects the hypothesis that the data come from a Poisson distribution. We should note that they are sensitive to different kinds of departure – individual frequencies that are too high or too low and overall spread, respectively. Thus, although the dispersion test will not be powerful for detecting aberrant frequencies, Pearson's chi-squared test (26.14) will be less good at detecting outlying individual values, which will typically have to be grouped in a neighboring cell.

The same method can be used to test for the goodness of fit to the other distributions we have introduced, remembering that for the normal distribution, for example, we need to deduct a further degree of freedom to allow for the estimation of the variance. The asymptotic chi-squared distribution of X_ν^2 results from approximation theory related to the Central Limit Theorem. An alternative approach also commonly encountered is based on asymptotic likelihood theory and involves a "deviance statistic" of the general form:

$$D = 2\sum_i [f_i \log(f_i/e_i) - (f_i - e_i)], \qquad (26.15)$$

reflecting individual contributions to the log-likelihood. This statistic also has an approximate chi-squared distribution when the expectations are not too small; the number of d.f. is the same as for X_ν^2. With moderately large amounts of information, it will give very similar results to Pearson's chi-squared. For Example 26.8, the value is 1.95, for which $P = 0.92$; in larger datasets, the two statistics will be closer for theoretical reasons. The deviance is frequently encountered for testing agreement in more complicated models, such as logistic and Poisson regressions (cf. chapter ▶Geographical Epidemiology of this handbook).

Example 26.9. Dispersion of traffic accident data
As a further example, we consider the annual numbers of road accidents in the UK town of Bolton, Lancashire, in the 10 years 1999–2008, posted on the Greater Manchester Transportation Unit website (http://www.gmtu.gov.uk/). The counts are shown in Table 26.3; there are too few to perform a goodness-of-fit test to the shape of each distribution. The dispersion test for the distribution of the 98 fatal accidents gives $X_9^2 = 12.4$ ($P = 0.19$), with 9 d.f., and this test is equivalent to the Pearson's chi-squared test (26.14) of the hypothesis that all the expectations e_i are equal to $98/10 = 9.8$. Substituting into the deviance statistic (Eq. 26.15) instead, we obtain $D = 12.9$, a value close to the dispersion statistic. It also has a chi-squared distribution with 9 d.f. Neither result is significant, but if we apply the same tests to the non-fatal serious accidents, we find a significant departure from the null hypothesis, with $X_9^2 = 17.81$, $D = 17.70$, $P = 0.037$, 0.039, respectively. This is not really surprising since traffic volumes and conditions will have changed quite a lot over 10 years, inducing different expectations for the different years. The fact that fatal accident rates appear not to depart from the null hypothesis is almost certainly due to the

Table 26.3 Annual numbers of road accidents in Bolton, Lancashire, by severity

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Slight	1,124	1,234	1,149	1,023	1,017	942	947	857	797	698
Serious	102	96	84	77	98	83	107	77	74	69
Fatal	4	15	11	10	7	14	7	14	7	9

smaller numbers. This is borne out by the dispersion statistic for the more numerous slight accident data, $X_9^2 = 253$, which is so extreme that it is not worth calculating the p-value.

26.4.7 Non-Parametric Tests

In Sect. 26.4.2 (Example 26.5), we tested the difference between two means using the fact that sample means are at least approximately normally distributed. This methodology assumes that we can ignore sampling variations in the estimation of the variances, which will be less true for small samples. The well-known t-distribution known as Student's t-distribution (see Sect. 26.5.1) provides methods allowing for this problem, but their reliability and effectiveness depend on the normality assumption. In small samples and with very non-normal parent distributions, these methods may give misleading p-values, and we seek different methods not dependent on assumptions about the form of the parent distribution. Tests of this kind are termed *non-parametric* tests.

Centrally important in such methods are those based on ranks. For the comparison of two independent samples, for example, the rank sum test ranks *all* the observations (in order of increasing magnitude) and then determines the sum W of the ranks in the smaller sample: a sum that is smaller than might be expected by chance is clearly indicative of a distribution located below that of the larger sample. Tables are available to determine the significance of W for given sample sizes, the null hypothesis being that the two samples come from the same parent distribution, which is the only assumption made. This test is usually known as the Mann-Whitney U-test, though it is also attributed to Wilcoxon (1945); variant tests based on ranks are available for the case of related samples (of paired observations) and for more complicated situations.

Example 26.10. Rank sum test of IQ depression data
Applying the method described above to the IQ depression data from Example 26.5 requires an adjustment for breaking the ties. The resulting sum of the ranks in the smaller sample is $W = 408.5$, giving $P = 0.058$ in a two-sided test, smaller than the p-value of 0.11 for the normal-based test in Example 26.5. One might have expected to find a smaller p-value in the procedure based on the normal distribution since it makes stronger distributional assumptions. In this example, however, the reliability and the power of the normal-theory procedure are impaired by the inflation of the variance estimate caused by the outliers – that is, the unusually small IQs – in both samples.

26.4.8 Monte Carlo Tests

The rank sum test is one of a class of tests known as *permutation tests*. The null hypothesis merely specifies that any of the values actually observed is equally likely to have come from either sample, subject only to the requirement that n_1 of the combined sample of $n_1 + n_2$ must come from the first sample. We can imagine writing the values on individual cards, shuffling the pack and selecting n_1 from the resulting pack to constitute an imitation first sample, the remaining ones constituting the second sample, of course. If we repeated this procedure a large number, say N, of times, we could compute a sequence of values $W_1, W_2, W_3, \ldots, W_N$, and this would provide an estimate of the true null distribution of W, which we could then use to check the calculated p-value for our observed statistic W_{obs}, say. Specifically, the proportion of the simulated Ws that are greater than or equal to (or less than or equal to) W_{obs} provides an unbiased estimate of the true p-value of the null distribution in a one-sided test of the right (or left)-hand tail of the distribution.

In fact, this check is unnecessary for the rank sum test since it is possible to calculate the null distribution of W for any given n_1, n_2. This is because the value of W depends only on the ranks, not the values of the actual observations. The idea does, however, suggest a very wide-ranging method for estimating the p-value for any test of a simple null hypothesis. We could, for example, test the difference between two independent samples using a suitable function of the actual observations, such as the z-statistic (Eq. 26.13) we have already used in Example 26.5; the permutation test allows us to estimate the p-value without having to specify any model for the parent distribution. Clearly, the precision of the estimate of the true p-value depends on N, and its s.e. is easily estimated using the binomial distribution. Tests of this kind are known as Monte Carlo tests.

Example 26.11. Monte Carlo test of IQ depression data
In an application of the Monte Carlo method to the IQ depression data (Example 26.5), 10,000 datasets were constructed by sampling from the 94 IQs observed in the two samples and selecting 15 IQs at random (and without replacement) for the depressed sample. Applying (26.13) to each dataset in turn yielded a set of z scores with mean -0.15 and variance 1.11, both substantially different from the expected theoretical values of 0 and 1. Of the simulated values, 322 were greater than or equal to the observed value $z_{\text{obs}} = 1.64$ so that a two-sided p-value is estimated to be 0.064. We might have expected the actual z-value to have the smaller p-value, but the latter is inflated by the large variances induced by the outlying observations. In any case, comparing individual p-values for different tests is not a good guide to the respective powers of the tests.

A variant of this procedure – also known as a Monte Carlo test – provides an exact test with a specified type I error (Marriott 1979). More generally, the term "Monte Carlo" includes a wide range of methods for statistical inference of considerable importance in modern statistics. None of them, however, is as simple and effective as the procedure described above for testing a simple null hypothesis.

26.5 Confidence Intervals

We saw in Sect. 26.3 that the sampling distribution of an estimator provides a way of quantifying how close an estimate is likely to be to the *estimand* or parameter it estimates. We now describe how to make this more precise by constructing *confidence limits* and *confidence intervals*.

26.5.1 Confidence Limits for the Normal Mean

Since we know that the mean of a sample from a normal distribution is itself normally distributed with mean μ and SD equal to $\sigma/\sqrt{n}$ (see Sect. 26.3.3), we can write, for example,

$$0.025 = \mathrm{P}[\bar{X} > \mu + 1.96\sigma/\sqrt{n}] = \mathrm{P}[\mu < \mu_L] \text{ where } \mu_L = \bar{X} - 1.96\sigma/\sqrt{n}. \tag{26.16}$$

The quantity μ_L so defined is known as a *lower 97.5% confidence limit* for μ and has the interpretation that if we calculate it from a large number of independent samples in the same manner, then it will be less than μ on 97.5% of occasions.

In the same way, we can calculate an upper limit, namely $\mu_U = \bar{X} + 1.96\sigma/\sqrt{n}$. Then we have $\mathrm{P}[\mu > \mu_U] = \mathrm{P}[\mu < \mu_L] = 0.025$, and so these two limits define a 95% *confidence interval*:

$$\mathrm{P}[\mu_L < \mu < \mu_U] = 0.95, \tag{26.17}$$

and the probability on the right-hand side of this equation is known as the *confidence coefficient* of the interval. This equation needs to be interpreted with care. At first sight, it looks like a prediction interval for a random variable called μ, giving the probability that μ lies between two fixed limits. But in the frequency approach to inference, this is quite wrong. We must interpret it instead as a probability statement about a random interval, giving the probability that it contains the unknown parameter μ. We must envisage hypothetical repetitions of the experiment in which some intervals do, and some do not, contain the unknown μ, as illustrated in Fig. 26.11. The confidence interval calculated on a given occasion tells us nothing else about the probability that this particular interval contains μ.

Example 26.12. Confidence limits for the HbA1c data
The non-obese OSA patients had a mean HbA1c of 5.4 and SD of 0.8 so, using the expressions above, it is straightforward to calculate a 95% confidence interval for the mean HbA1c in the non-obese OSA patients as $5.4 \pm 1.96 \times 0.8/\sqrt{26} = (5.09, 5.71)$. Note that we have used the SD for the sample because we are no longer making the assumption that the sample was drawn from the standard population of Telecom subjects.

Because the sample is quite small and we have used an estimate of the underlying SD, which is itself subject to error, we can improve our calculation. It can be shown that, for a sample of size n from the normal distribution with parameters μ and σ, the statistic $T = \sqrt{n}(\bar{X} - \mu)/s$ has the distribution known as the Student's t-distribution[12] with $n - 1$ degrees of freedom (d.f.) – a parameter reflecting the number of observations contributing to the estimate of the variance – *regardless of the value of* σ. This remarkable result can be used to refine many of the procedures involving small samples from the normal distribution. It is one of the few really

[12] Student was the pen name of W.S. Gosset (1876–1937), who worked for Guinness, by whom he was not permitted to publish the results of his work.

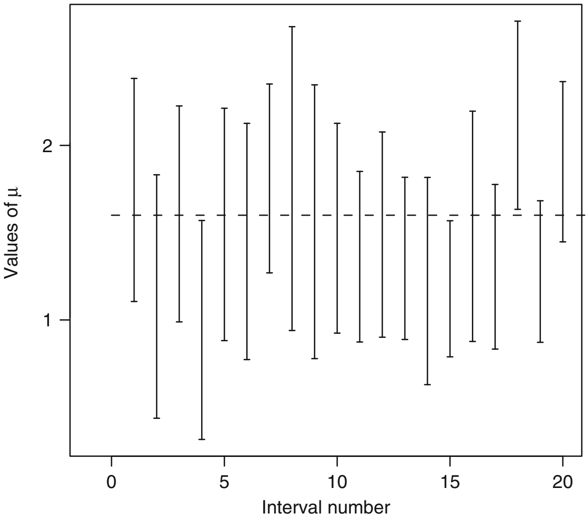


Fig. 26.11 Hypothetical repeated calculations of a 95% confidence interval for μ, which is shown by the *dotted line* since its precise position is unknown. Intervals 4, 15, and 18 just fail to include the true μ, compared with one interval expected

elegant and exact results in statistics and is unique to the normal parent distribution. As might be expected, the t-distribution looks quite like the normal distribution if n is not too small and indeed it gets closer as the sample size increases. For $n = 26$, the 97.5% upper limit of the appropriate t-distribution (i.e., with 25 d.f.) is 2.06, around 5% larger than 1.96. A recalculation of the confidence interval in Example 26.12 therefore gives $5.4 \pm 2.06 \times 0.8/\sqrt{26} = (5.08, 5.72)$, which is also about 5% longer than the approximate interval based on the normal distribution. We note that this interval does not include the reference (Telecom) value of 5.05, and it is quite easy to see from the arithmetic that this means that we would have rejected the null hypothesis $H_0 : \mu = 5.05,\ \sigma = 0.8$ in a two-sided t-test at the 5% level.

26.5.2 Confidence Intervals and Hypothesis Tests

This observation – that we would reject a null hypothesis concerning a parameter that does not fall in the confidence interval – is perfectly general, and it affords a powerful approach to the construction of confidence intervals. We have the following principle:

> ***Equivalence Principle:*** *Given a test of a simple null hypothesis $H_0 : \mu = \mu_0$, with type I error α, the set of all values of μ_0 that would not be rejected by this test defines a $1-\alpha$ confidence interval for μ.*

Clearly, if we construct a $1 - \alpha$ confidence interval in this manner from a particular test statistic and the value μ_0 is not included in this interval, then the test will reject $H_0 : \mu = \mu_0$ at level α. We assume that the resulting set of parameters forms a single interval; otherwise we can use the word "region." This word also covers the case where we have more than one parameter and the resulting region is in two or more dimensions. It is frequently much easier to carry out a test of a simple parametric hypothesis than to calculate the corresponding confidence interval or region; in this case, the limits of such an interval can often be conveniently calculated by considering a number of hypothesis tests. An example follows in the following subsection.

26.5.3 Confidence Limits for Binomial Data

As with hypothesis testing, discrete data pose a problem. We consider, for example, a confidence interval for the parameter p of the binomial distribution. As with the hypothesis testing situation, we could use a normal approximation to the distribution of X with a SD of $\sqrt{mp(1-p)}$. But to substitute $\hat{p}$ for p is unsatisfactory for constructing a confidence interval since we cannot assume a single value for p – it is after all our object to estimate this unknown parameter. However, if we can find values (p_L, p_U) of p such that

$$\frac{x - mp_L - \frac{1}{2}}{\sqrt{mp_L(1-p_L)}} = 1.96 \text{ and } \frac{x - mp_U + \frac{1}{2}}{\sqrt{mp_U(1-p_U)}} = -1.96, \qquad (26.18)$$

then these values of p will be 95% confidence limits for p to a good approximation provided mp and $m(1-p)$ are not too small – say, at least 10. As before, the $\frac{1}{2}$s are a continuity correction that can substantially improve the approximation.

For smaller values of mp and $m(1-p)$, however, we should like to find a method analogous to the exact test described in Sect. 26.4.5. Now we must find the critical values of p that just make the observation x significant in the left- or right-hand tails, that is, we seek solutions p_L, p_U to the equations:

$$\sum_{j=0}^{x} \binom{m}{j} p_U^j (1-p_U)^{m-j} = \sum_{j=x}^{m} \binom{m}{j} p_L^j (1-p_L)^{m-j} = \alpha/2. \qquad (26.19)$$

Evaluating the terms in these sums for a given p and solution by trial and error is feasible, though not very elegant. In fact, an analytical solution exists that enables the answer to be evaluated using available computer routines or statistical tables; the reader is referred for a simple account to Armitage et al. (2002, Sect. 4.6). This so-called "exact" method is not quite as perfect as it sounds since it is actually conservative: the *coverage probability* or probability that the interval really does cover the true value of the parameter is guaranteed to be at least $1-\alpha$ but may in fact be appreciably higher.

Example 26.13. Confidence interval for probabilities in the meningitis data
If we apply the normal approximation (26.18) to the proportion discussed in Example 26.1, Sect. 26.2.1, we find a 95% confidence interval of (0.489,0.760). Multiplying the interval by $m = 52$ gives (25.4,39.5) for the mean of the binomial distribution, which is therefore almost certainly greater than 10; this suggests that the normal approximation should be

satisfactory. The conservative “exact” method, for comparison, gives (0.490,0.763), which is in good agreement. The fact that the interval just includes 0.5 accords with the finding that we cannot quite reject H_0.

Neither interval is truly exact, of course. The actual coverage error depends on the true value of p, and numerical calculations for different values of p lead to the conclusion that the errors in both intervals can be considerable and vary rapidly with changing values of p. Much recent work has been concerned with identifying the best method to use in given circumstances. For example, in a recent review, Agresti and Gottard (2007) recommend a method based on the “mid-P,” a modification to the exact test in which only half the probability of the observed x is included in the tail probability.

Example 26.14. Survival probability in meningitis data
A more challenging example is provided by data on the survival of the children with meningitis. Those with platelet counts below 10^{11} per liter have a markedly poorer survival, only 4 of the 21 in this category surviving the disease. Applying the normal approximation to this gives a 95% confidence interval for p of (0.063,0.426); 21 times the lower limit is only 1.32, which indicates that the normal approximation is unlikely to be good in this case. Application of the exact conservative method yields (0.054,0.419), and this method should probably be regarded as preferable in this case.

26.5.4 Confidence Limits for Poisson-Distributed Data

As with the binomial distribution, use of the normal distribution with the estimated standard error of the mean λ can be unsatisfactory for the construction of confidence intervals since this estimate is itself subject to significant sampling error; in any case, the true value that would be relevant is the value near the upper or lower limit we are trying to construct. It is not recommended that this approach should be used for the Poisson distribution unless the mean of the data exceeds 100 (see Armitage et al. 2002, Sect. 26.5.2). However, we can use the method analogous to (26.18), which gives us the following quadratic equations for the 95% limits, (λ_L, λ_U), based on a single observation x:

$$\frac{x - \lambda_L - \frac{1}{2}}{\sqrt{\lambda_L}} = 1.96 \text{ and } \frac{x - \lambda_U + \frac{1}{2}}{\sqrt{\lambda_U}} = -1.96. \qquad (26.20)$$

The exact method also proceeds along the same lines as in the binomial case (26.19), but using the probabilities for the Poisson distribution; this is to be preferred for inference from small observations, although, as with the binomial, it can be very conservative. Further details of this method are given by Armitage et al. (2002), Sect. 26.5.2.

Example 26.15. Daily rate of emergency admissions
We apply the different methods to the emergency admissions data from Example 26.3, remembering that for our purposes the data are equivalent to a single observation equal to the total number of admissions, 408. To obtain an interval for the mean number of admissions per day, we merely divide by 104, the number of observations. Because the

total is quite large, we would expect the methods to give very similar results, and this is indeed the case: (26.20) is perhaps the method of choice here and gives a 95% confidence interval of (3.56,4.33) for the daily mean. The rather cruder interval $\hat{\lambda} \pm 1.96\sqrt{(}\hat{\lambda})$ gives (3.54,4.30), while the exact method, choosing the values of λ to give tail probabilities of 2.5% (or referring to Eq. 5.6 in Armitage et al. (2002)), gives (3.55,4.32).

26.5.5 Perspective on Confidence Intervals

As mentioned above, we can construct confidence regions in multiparameter situations, to give, for example, an elliptical region for a pair of parameters. As might be expected, the methodology is distinctly more difficult, since we are no longer able to invert a simple inequality as we did in Eq. 26.16; only in a few cases is there any analytical theory available to help us, and we often have to resort to serious amounts of computational trial and error. An even more challenging problem arises in multiparameter situations when we try to construct an interval for a single parameter. Difficulties arise because parameters are generally interrelated – we should avoid speaking of their independence in a frequentist context since this is a property of random variables, and their probability distributions. But more frequently than not, parameter estimates are themselves dependent variables and this means that information on one parameter conveys information (albeit fairly indirectly) on other parameters. It is therefore not generally permissible merely to substitute estimates of the uninteresting, so-called *nuisance* parameters and construct an interval using methods that assume we know the other parameters. It is true that this is what we have already done in this chapter when we used a large-sample estimate for σ in a normally distributed sample, for example – but there are special considerations that apply to the normal distribution. We should be aware that this does increase the uncertainty in our inferential process somewhat and avoid doing it when the sample sizes are small. Fortunately, there are approximate methods available for this problem through the use of the likelihood (Sect. 26.6).

It is generally agreed that constructing a confidence interval or region is more valuable and informative than the execution of a single significance test, not least because, as the equivalence principle discussed above shows, every confidence interval incorporates many hypothesis tests of interest, whereas a single hypothesis test does not permit construction of an interval. However, the same principle makes it clear that the difference is not of a fundamental kind. The confidence interval still describes long-run properties of a procedure using probabilities on the experimental scale, *not* on the scale of plausibility of our model. It may even be that, as with the p-value, the apparent quantification of uncertainty can give an impression that is more misleading than a formal significance test. Assessing the likeliness of a parameter value still requires a similar inferential inversion: we could, for example, conclude that it is unlikely that μ lies outside the calculated limits only because, if it does, something relatively unlikely has occurred and we know this does not happen very often. It is very important to keep this limitation in mind in all forms of frequentist inference.

26.6 Likelihood-Based Inference

We return now to the use of the likelihood, which – as we have already seen – provides a good general method for constructing point estimates. We have also intimated that the further a parameter value is from the MLE, the less plausible it is as a contending value, and we can quantify this statement by measuring plausibility in terms of the ratios of the likelihood L. Thus, we can construct a *likelihood interval* for a parameter μ containing all the values of μ for which $L(\mu) > \rho L(\hat{\mu})$, where ρ is a fraction such as 0.25, which we may call a *plausibility fraction*. We would then declare that values within the interval are at least a quarter as likely as the most likely estimate, where "likely" refers to plausibility in terms of the probability of the observed results for given values of μ. The value of ρ remains arbitrary, and it is not obvious that there is anything "universal" about our chosen value – that is, that it would make as much sense in one statistical context as another.

There is, however, a close connection with the frequentist concept of a confidence interval. The Neyman-Pearson lemma referred to in Sect. 26.4.3 determines a best test criterion in terms of the ratio of the likelihoods under two hypotheses, and consideration of the equivalence principle suggests that – as well as being a good way of establishing a test statistic and critical region for formal testing purposes – the ratio also provides a sensible way of establishing a confidence interval. Thus, we are led once more to the choice of an interval based on values of the likelihood, but now we will hope to choose the critical ratio ρ in such a way that it corresponds to a known confidence coefficient. We try this out for the normal mean, continuing Example 26.2.

26.6.1 Likelihood Interval for the Normal Mean

Working with the log of the likelihood and remembering that we are assuming σ to be known, we easily find from (26.8) that

$$\Lambda = \log(L(\hat{\mu})/L(\mu)) = \ell(\hat{\mu}) - \ell(\mu) = n(\bar{x}-\mu)^2/2\sigma^2 \tag{26.21}$$

since $\hat{\mu} = \bar{x}$. Hence, a region given by $L(\hat{\mu})/L(\mu) < 1/\rho$ is equivalent to one determined by

$$n(\bar{x}-\mu)^2/\sigma^2 < -2\log(\rho), \text{ or equivalently } \bar{x} - z_{\alpha/2}\,\sigma/\sqrt{n} < \mu < \bar{x} + z_{\alpha/2}\,\sigma/\sqrt{n}, \tag{26.22}$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution for some probability α. Thus, the likelihood interval is equivalent to the interval we have already constructed in Eq. 26.17. Moreover, we can now relate the confidence coefficient $1-\alpha$ to the plausibility fraction ρ through the equations

$$\rho = \exp(-\tfrac{1}{2}z_{\alpha/2}^2) = 0.147 \text{ in the case } \alpha = 0.05. \tag{26.23}$$

Thus, a 95% confidence interval corresponds to the values of μ which are at least $\rho = 14.7\%$ as likely as the MLE. Our original proposal of $\rho = 0.25$ corresponds to a confidence coefficient of $1 - \alpha = 0.904$, just over 90%.

26.6.2 Likelihood-Ratio Tests

The same ideas apply of course to hypothesis testing – as we would expect in view of the equivalence principle (Sect. 26.5.2). If we can find the distribution of the likelihood ratio or its logarithm under a given H_0, then it will, in many circumstances, provide a good test. This will apply even when the ratio cannot be expressed in terms of a simple statistic and when H_0 specifies several parameters. Very often the distribution theory will not permit us to compute the distribution of the likelihood ratio exactly, but there is a very useful approximation theorem available.

Suppose that we have a simple null hypothesis $H_0 : \theta = \theta_0$, where θ is a parameter that may have several – say k – elements. Then, under fairly general conditions, and for reasonably large samples, the quantity $\Lambda = 2 \log \left(L(\hat{\theta})/L(\theta_0) \right) = 2(\ell(\hat{\theta}) - \ell(\theta_0))$ has a distribution that is approximately the chi-squared distribution with k degrees of freedom almost regardless of the true underlying distribution of the data. The approximation improves as the sample size increases, and it becomes useful for samples of moderate size in many situations. Moreover, the same result holds when H_0 is a composite hypothesis, requiring the estimation of $p < k$ parameters, in which case Λ has an approximately chi-squared distribution with p d.f. provided $L(\theta_0)$ is taken to be the maximum value of L consistent with H_0. This is one of a number of "asymptotic" results of considerable importance in statistics.

L-R Test for Difference of Two Proportions From (26.6), we find that the log-likelihood for binomial data is given by

$$\ell(p, m; x) = C + x \log(p) + (m - x) \log(1 - p), \tag{26.24}$$

where C is a constant which will cancel when we form the log-likelihood ratio. The log-likelihood for the two binomial samples is simply $\ell(p_1, m_1; x_1) + \ell(p_2, m_2; x_2)$, and the maximum value of this is obtained by substituting $\hat{p}_i = x_i/m_i$ for $i = 1, 2$. Under $H_0 : p_1 = p_2$, however, both ps are estimated by the same parameter, and it can be seen that this must be $\hat{p} = (x_1 + x_2)/(m_1 + m_2)$. Substituting these values into the expression for the binomial log-likelihood (26.24) gives the statistic Λ:

$$\Lambda = 2(\ell(\hat{p}_1, m_1; x_1) + \ell(\hat{p}_2, m_2; x_2) - \ell(\hat{p}, m_1 + m_2; x_1 + x_2)), \tag{26.25}$$

which has a χ^2_1-distribution.

Example 26.16. Comparison of survival rates in children with meningitis
Bortolussi et al. (1978) also record the survival of the children with meningitis according to their birthweight: of 26 children who were underweight (<2,500 g) at birth, 17 died, compared with 8 out of 26 with normal weights. Substituting these figures into Eq. 26.25 gives $\Lambda = 6.37$, which is significant at the 5% level ($P = 0.0116$).

26.6.3 Likelihood-Based Regions

The same theoretical result, in conjunction with the equivalence principle, provides a general method for constructing an approximate interval when there is no exact distribution theory, and also for confidence regions when there are two or more parameters. For example, Fig. 26.12 shows an approximate likelihood-based confidence region for the parameters (μ, σ) in the HbA1c data formed by the intersection of the log-likelihood surface and the plane $\ell(\mu, \sigma) = \ell(\hat{\mu}, \hat{\sigma}) - 3.00$, the constant 3.00 being half the 95% percentile of the chi-squared distribution with 2 d.f., that is, $\mathrm{P}[\chi^2_2 > 5.99] = 0.95$.

26.6.4 The Likelihood Approach

We have been at pains to emphasize that the probabilities derived in the frequentist inferential framework represent the long-run performance of statistical procedures, not the uncertainty in individual cases. Although this long-run performance is an invaluable guide to the choice of procedure, it leaves many statisticians unhappy and keen to find an alternative way to quantify the individual uncertainties. The support for different values of a parameter offered by the likelihood is undoubtedly

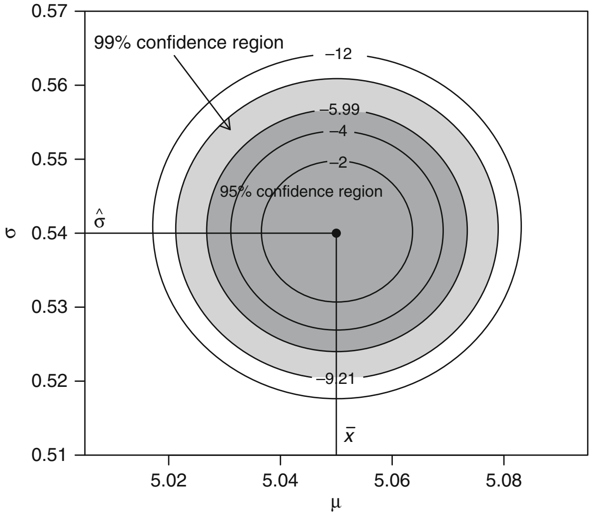


Fig. 26.12 Contours of the log-likelihood showing 95% and 99% confidence regions for μ and σ from the HbA1c data ($n = 3,240$; $\bar{x} = 5.05$; $\hat{\sigma} = 0.54$)

meaningful, and some authors[13] speak of the "supported range" for a parameter, meaning the range of values at least ρ times as likely as the maximum. We are now in a position to answer the question posed at the beginning of this section – namely whether the value of ρ we use would have the same significance in different situations. In terms of the coverage probability, the answer would seem to be "yes," but only provided we have the same number of parameters in the different situations. For example, a supported range with $\rho = 0.147$ corresponding to 95% coverage with one parameter should be compared with a *supported region* with $\rho = 0.05$ when there are two parameters, as in Fig. 26.12. In words, this means that a two-dimensional 95% confidence region corresponds approximately to those parameter values for which the likelihood is at least 5% of the maximum value.[14]

26.7 The Bayesian Approach

We saw in Sect. 26.2.2 that Bayes' theorem can be used to derive a probability for the truth of some statement about our model, for example, that a parameter lies in a given interval. We also saw that the posterior probabilities yielded by a Bayesian analysis depend on the prior probabilities that represent our scientific preconceptions.

We illustrate the ideas by revisiting our former examples, starting with the binomial distribution. We first introduce a notation to help to distinguish the different distributions involved. When necessary, we will continue to use forms of the symbol f to denote densities of observations, as in (26.2), for example. Because parameters commonly encountered are real numbers as opposed to integers, we also need to describe their probability distributions by means of density functions; to avoid confusion, we will use the symbol g instead of f for the densities of the parameters. We will distinguish between the prior and posterior densities for a parameter θ by using the symbols $g_0(\theta)$ and $g_1(\theta)$, respectively. The relationship between these follows from Bayes' theorem in the general form:

$$g_1(\theta|x) \propto L(\theta; x) \times g_0(\theta), \qquad (26.26)$$

that is, the posterior density (of the parameter) is proportional to the likelihood times the prior density. It will be seen that the likelihood is the factor that introduces the information in the data to modify our perception of where θ is likely to lie. It is generally not necessary to specify the proportionality factor since this can

[13] In an epidemiological context, see, for example, the book by Clayton and Hills (1993).

[14] It is an interesting mathematical property of the χ^2_2-distribution that this value of ρ is exact, and indeed, the plausibility fraction for a two-dimensional $1-\alpha$ confidence region is α for *any* α, at least to the extent that 2Λ is approximately distributed as χ^2_2.

be obtained by requiring that the posterior density integrates to one (i.e., has area underneath it equal to one).

26.7.1 Bayesian Estimation of the Binomial Parameter

Suppose we have observed x positive responses out of m trials in a binomial experiment and that we wish to estimate the probability p of a positive response. Applying (26.26) to the parameter p, we have

$$g_1(p;x) \propto L(p;x) \times g_0(p), \tag{26.27}$$

where $L(p;x)$ is the binomial likelihood, as defined in Eq. 26.6, and $g_0(p)$ is the supposed density of the parameter p prior to the observation of any data.

A major issue is the question of how to choose the prior density given that, by its very nature, it will usually be subjective. We will need a family of prior densities defined on the interval [0, 1], and this is provided by the so-called beta distribution family, with density:

$$g_0(p;a,b) = p^{a-1}(1-p)^{b-1}/\mathrm{B}(a,b)\ ,\ 0 \le p \le 1,\ a,b > 0, \tag{26.28}$$

where a and b are parameters taking any real values greater than zero. (We need not be concerned with the proportionality factor $\mathrm{B}(a,b)$, which involves factorial-like quantities depending only on a and b.) The mean and variance of this distribution are $a/(a+b+1)$ and $ab/(a+b)^2(a+b+1)$, while its *mode* (or the value of p with the largest density) is at $(a-1)/(a+b-2)$, provided $a,b > 1$. Figure 26.13 shows different members of the family as a and b vary. If, for example, we thought in advance of the observation that the most plausible value for p was 0.25, we could ensure that the distribution has this value for the mode by choosing $b = 3a - 2$. If we then choose $a = 2$, $b = 4$ we will have a SD of $\sigma = 0.178$. The spread of this distribution represents our view of how close the true value is likely to be to its center; if we were more confident that 0.25 was the most likely value for p, we might choose $a = 6$, $b = 16$, giving $\sigma = 0.093$, about half the previous value. If we really had no idea of an appropriate value before seeing the data, we could opt for the *rectangular distribution*, which has constant density and assigns to all intervals of a given length the same prior probability of containing p. Such a choice gives what is known as an *uninformative prior*, for obvious reasons.

Any (positive) values of a and b can be substituted into (26.28), giving for the posterior density

$$g_1(p) = p^{a+x-1}(1-p)^{b+m-x-1}/\mathrm{B}(a+x, b+m-x), \tag{26.29}$$

which is another beta distribution but with revised parameters. Comparison with Eq. 26.28 enables us to see that the posterior mean and mode are given by

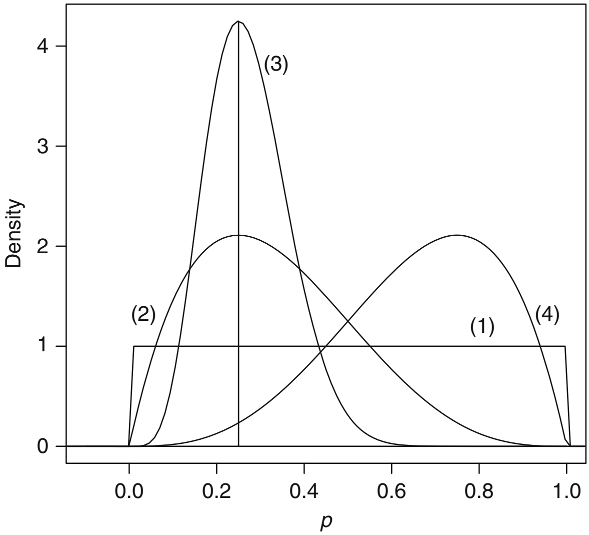


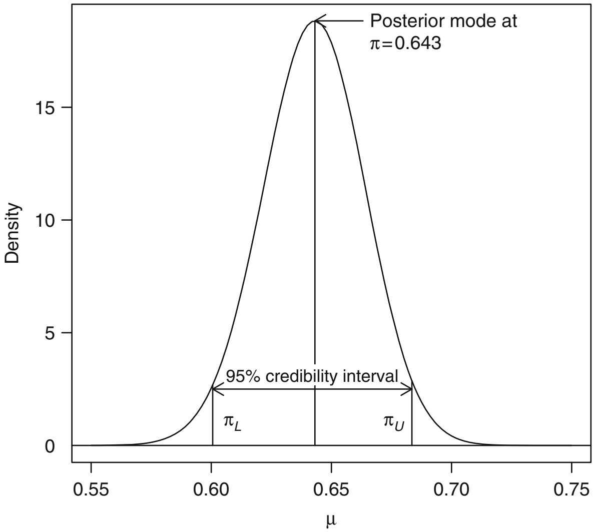
Fig. 26.13 Density of the beta distribution for various values of (a, b): (1) (1,1) – rectangular distribution; (2) (2,4) – see text; (3) (6,16) – same mode; (4) (4,2) – note the symmetry of distributions (2) and (4)

$$\text{mean: } (a+x)/(a+b+m+1) \text{ and mode: } (a+x-1)/(a+b+m-2), \text{ respectively.} \quad (26.30)$$

With binomial data, the prior and posterior densities have the same form only for the beta family of densities; a prior family with this attractive feature is called a *conjugate prior*. It is clear that the posterior distribution has the form obtained by increasing a and b by x and $m - x$, respectively. This suggests that the information in a beta prior distribution with parameters a and b can be compared with that in a binomial experiment in which a successes are observed in $a + b$ trials. Hence, if, prior to our study, we only have information from a previous study with this outcome, we might think it rational to choose our prior with these parameters. Furthermore, we may, even in the absence of such information, represent our beliefs in terms equivalent to hypothetical observed data: thus, we may represent our belief about a binomial probability by imagining an experiment in which we have observed a positive outcomes in $a+b$ trials. This *information equivalence* is a helpful way of thinking about the synthesis of prior and current information and is made possible by the use of the conjugate prior.

Example 26.17. Bayesian estimation of the sex ratio in the meningitis data
Continuing Example 26.13, we need to fix a prior distribution for p, the probability of a child being male. In fact, it is now well known that male children are far more susceptible to infections than females, though the reasons for this is are undoubtedly complicated and still rather obscure. A decade before Bortolussi et al.'s (1978) study (which was not primarily concerned with the sex ratio), Washburn et al. (1965) had reported a very large search of the literature on the sex ratio, covering many common infections. For meningitis in newborn infants (i.e., under 1 month old), the search revealed 296 males out of 460 children with meningitis (64.3%). The information equivalence concept suggests that we might reasonably choose a prior distribution for p having $a = 296$, $b = 460 - 296 = 164$ to represent the previously available information. We could then combine this with the

Fig. 26.14 Posterior density of μ for meningitis data, showing the symmetric 95% credibility interval



Bortolussi data of 33 males out of 52 children to give a posterior distribution with mean $(33 + 296)/(52 + 460 + 1) = 0.641$ and SD 0.021, though this would only make sense if we believed that the experimental contexts were very similar. It will be seen that this distribution is heavily dominated by the large amount of prior information in the Washburn study; it is shown in Fig. 26.14, with a 95% *credibility interval*, that is, an interval encompassing 95% of the posterior probability. The latter is sometimes called a *Bayesian confidence interval*, but the former term is perhaps to be preferred, in order to avoid conceptual confusion. There are of course many ways in which the interval could be chosen: one might, for example, wish to ensure that the probability of passing either limit is the same, say $\alpha/2 = 2.5\%$; alternatively, we might want the shortest possible interval, which is achieved by taking the *highest posterior density (HPD) interval*, containing just those points p whose density is above an appropriate value. Figure 26.14 shows the first of these methods, which is significantly easier to compute and returns a credibility interval of (0.601,0.684). In this case, the 95% HPD limits are the same to three significant figures, though the two-tail probabilities are respectively 2.4% and 2.6%, that is, it makes little difference which method we choose when the distribution is reasonably symmetric. The most “likely value” of p is at the posterior mode, which is 0.643, only slightly larger than the posterior mean; the choice between mean and mode to summarize the data depends on how we wish to interpret the results. We observe that the interval is very considerably shorter than the confidence interval of (0.489,0.760) given in Example 26.13. This is clearly due to the extra information imported from the earlier study.

26.7.2 Bayesian Estimation of the Normal Mean

For the normal distribution, we will for convenience consider a single observation x, though this will not, of course, preclude us from taking x to be a more complicated statistic, such as a sample mean. We continue to focus on inference for the mean, making the assumption that we know the variance σ_x^2 of x well enough to be able to consider it known; this will be reasonable for inference on μ from moderately large samples. From the structure of the likelihood (26.7) – that is, from the quadratic term in the exponent – we can see that the conjugate prior is itself normal, so we suppose that μ is normally distributed with, say, mean τ_0 and variance κ_0^2. The resulting probability densities are easily manipulated to give the posterior density as another normal distribution, its mean and variance being respectively:

$$\tau_1 = \frac{x/\sigma_x^2 + \tau_0/\kappa_0^2}{1/\sigma_x^2 + 1/\kappa_0^2}\,; \qquad \kappa_1^2 = \frac{\sigma_x^2 \times \kappa_0^2}{\sigma_x^2 + \kappa_0^2} = \left(\frac{1}{\sigma_x^2} + \frac{1}{\kappa_0^2}\right)^{-1}. \tag{26.31}$$

It will be seen that the posterior mean is a weighted average of the observed value x and the prior mean τ_0 using the reciprocals of the variances as weights. This is very reasonable since the information we have from each of the two sources is well described by the inverse of the variance. A similar view of information as the reciprocal of variance enables us to see that the information in the posterior distribution $(1/\kappa_1^2)$ is the sum of that in the prior distribution and that from the observation.

As with the binomial distribution, we could consider the parameters of the prior distribution as representing prior information equivalent to a previous experiment, in which we have an observation from a normal distribution with parameters τ_0, κ_0, remembering that our single observation may well be the mean of a sample with standard error κ_0. The formulae in (26.31) then merely represent a pooling of the information in the two samples. However, using a large previous experiment to provide prior information is likely to give it overwhelming weight, and this only makes sense if we believe that the population in that experiment has the same mean as the one we are trying to estimate from our present data, though, in this case, we would hardly need the present data to provide new information. In practice, it is most unlikely that a population studied previously is completely relevant to our present context, and so we must then consider the question of how different the respective populations might be, a question which of course has nothing to do with the precision implied by the small sampling error in a big experiment.

Example 26.18. Bayesian estimation of the mean HbA1c for OSA data
Choosing the prior distribution in this example is distinctly challenging. If we use the Telecom study and invoke the principle of information equivalence discussed above, we obtain a prior distribution with mean and SD given by:

$$\tau_1 = \frac{5.4 \times 26/0.8^2 + 5.05 \times 3,240/0.54^2}{26/0.8^2 + 3,240/0.54^2} = 5.051\,;$$

$$\kappa_1 = \left(\frac{26}{0.8^2} + \frac{3,240}{0.54^2}\right)^{-\frac{1}{2}} = 0.0095. \tag{26.32}$$

The posterior distribution is heavily influenced by the Telecom data, which we have implicitly assumed to reflect a population with the same mean HbA1c as in the OSA study. In the light of the very highly significant difference between the two means, we know this

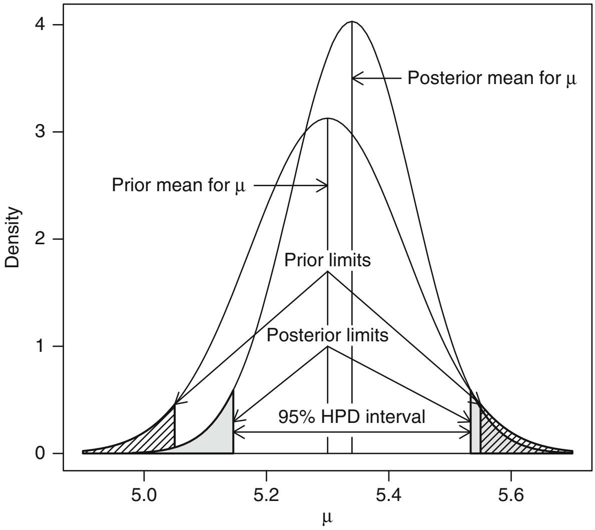


Fig. 26.15 Prior and posterior densities for the HbA1c – OSA data chosen by the method of prior limits, showing the resulting 95% credibility interval

is almost certainly quite wrong. For one thing, it is known that HbA1c depends on such factors as age and sex and we should ideally take account of this information. Moreover, the Telecom study is a sample of workers who will almost certainly be younger and healthier than the patients in the OSA sample. A scientist familiar with the nature of the pathologies involved would anticipate a connection of OSA with high HbA1c levels since both are associated with overweight, though with no other information to go on it would be difficult to know the extent of the difference. This illustrates one of the difficulties of a simple Bayesian analysis: the process requires us to choose the prior distribution in advance of seeing the current data, yet the latter may tell us that the assumptions on which the choice of prior is based are quite mistaken. We will suppose, for the sake of our example, that a more reasonable prior distribution is chosen, with mean $\tau = 5.30$ and SD $\kappa = 0.125$. We will discuss below (Sect. 26.7.4) how this might have been arrived at. The resulting posterior distribution now has mean 5.34 and SD 0.098 and is illustrated in Fig. 26.15. It gives a 95% HPD credibility interval (which is also of course symmetric) of (5.15,5.53).

26.7.3 Choosing the Prior

The problem of prior choice is central to a Bayesian analysis. Some workers are quite happy to use an uninformative prior, but there are serious difficulties with this. For one thing, in most cases (including the normal), the range over which a parameter must be specified is infinite, and it is not possible to specify a constant density over such a range that is *proper*, that is, that integrates to one. This turns out not to matter too much: it merely means that the density of the parameter, being constant, does not appear on the right-hand side of equations like (26.27) and this does not prevent us from finding the proportionality constant that delivers a proper posterior density. In this case, the density is just a scaled version of the likelihood so that the HPD credibility interval is then just a likelihood interval; indeed, in the case of the normal distribution with known σ^2, it is also equivalent to the frequentist confidence interval, reflecting the fact that no prior information has been provided.

A more serious difficulty is that uninformative priors are not invariant under changes of scale of the parameter. Suppose, for example, that we wanted to specify that the density of the logit[15] of a probability parameter p was constant, we would require a density that is quite different from the rectangular density we obtain by putting $a = b = 1$ in Eq. 26.28; needless to say, it would lead to different results. The implication is that there is no such thing as a truly uninformative prior distribution: by choosing a constant density function, we are implicitly choosing a scale on which we believe the density is constant, and this conveys information in itself.

In any case, it is usually quite unreasonable to claim that all values of a parameter are equally likely. Nobody could believe that a diastolic blood pressure, for example, is as likely to lie between 1,000 and 1,001 mm Hg as between 90 and 91 mm, since it is certain that a patient could not survive such a high blood pressure. Dependence on uninformative priors is forcefully attacked on these grounds by Greenland (2006) in an important paper advocating Bayesian methods. His solution is to use what we might call *prior limits*.[16]

26.7.4 Prior Limits

Given that some of the most useful outcomes of a Bayesian analysis will be statements like $\mathrm{P}[\mu > \mu_U] = 0.025$, we observe first that the general statement of Bayes' theorem, Eq. 26.4, permits us to consider such statements about our model and so in principle to derive the posterior version of this probability from the prior version. Unfortunately, we cannot do this without specifying the whole prior distribution and integrating over it in a calculation that will not generally be straightforward.

However, once we have specified a prior distribution family with, say, k parameters, we can in principle determine those parameter values by specifying k limits like that for μ specified above. We can achieve this, for example, using an upper and a lower limit with either a beta or a normal prior. It has to be admitted that only in the case of the normal distribution is the determination of the corresponding parameters really straightforward: the beta prior, for example, requires us to be able to manipulate the quantiles of the beta distribution, which

[15] logit(p) is defined as $\log(p/(1-p))$.

[16] In spite of the usefulness of this approach, this expression does not appear regularly in the literature as yet.

may put off those without a facility for mathematics and access to a good computing system.[17] Part of Greenland's argument, however, is that – just as in the frequentist approach – Bayesian analyses can be carried out using normal approximations, and the comparative simplicity of the calculation for the normal distribution makes this an attractive general method: he exemplified it with an example for the log of the odds ratio based on its asymptotic normal distribution.

For the normal distribution, we could, for example, choose upper and lower prior limits μ_{0U}, μ_{0L} such that we believe that

$$\mathrm{P}[\mu > \mu_{0U}] = \mathrm{P}[\mu < \mu_{0L}] = 0.025. \tag{26.33}$$

We can then find the mean and variance of the prior distribution from the equations

$$\tau_0 = \frac{1}{2}(\mu_{0U} + \mu_{0L})\ ; \quad \kappa_0 = \frac{\mu_{0U} - \mu_{0L}}{2 \times 1.96} \sim \frac{1}{4}(\mu_{0U} - \mu_{0L}). \tag{26.34}$$

Hence, we have easy formulae for determining the parameters τ_0, κ_0 of the prior distribution from two limits, which we still need to specify, of course. In many ways, however, choosing such limits is a more natural subjective exercise than choosing the mean and standard deviation of the prior density. We can invoke the language of betting to answer questions such as "What are the odds that the mean of the blood pressure lies above μ_{0U}?"; we emphasize that these odds refer to our probabilistic beliefs about where the true population mean lies, not to where we might expect the mean of a sample to lie. The answers to any two such questions determine our prior normal distribution; if we can invert the question to arrive at odds of 39:1, we can use the simple formulae of (26.34).

Example 26.19. Prior limits for HbA1c data
Continuing Example 26.18, we might suppose that an investigator is aware of the Telecom study, but also of the fact that OSA patients are likely to have raised HbA1c levels – he may, for example, have seen a number of cases where the level was higher than 6.0. He will also be aware that HbA1c assay standards have changed since the Telecom study was carried out and more typical values for HbA1c in the adult population have $\mu = 5.25, \sigma = 0.70$ (see, e.g., Myint et al. 2007). He will perhaps regard the Telecom study as representing a floor below which it is highly unlikely that his patients will record values; he might therefore choose the Telecom mean of 5.05 as his 2.5% lower limit. He might think that the few cases he has seen are high partly by chance and choose 5.55 as his 2.5% upper limit (always remembering that this is a limit for how high the mean probably is, not the percentile of the population). In this case, his (normal) prior distribution will have a mean of $\tau = 5.30$ and a standard deviation of $\kappa = 0.125$ (using the approximating denominator of 4 in (26.34)). Substituting into (26.31) and remembering that the quantity x stands for the sample mean $\bar{x}$ with estimated variance s^2/n, we obtain the limits illustrated in Fig. 26.15.

[17] It is relatively easy to write a function performing these calculations in R, for example.

26.7.5 Bayesian Hypothesis Testing

Hypothesis testing is a less common component of Bayesian analyses, which typically tend to place the emphasis on the estimation of parameters. This is partly because such analyses are invariably formulated in terms of parametric models, for which hypotheses can be represented as statements about the values of the parameters. The probability that a particular hypothesis is true can thus be derived from the (prior or posterior) distribution of the parameter. For example, the probability that a certain composite one-parameter hypothesis is true can be represented as the probability that the parameter takes particular values.

For simple hypotheses, however, a typical conjugate prior density for a parameter will assign the probability zero to a specific parametric hypothesis such as $p = 0.5$. This is a warning that it is rarely reasonable to regard a simple hypothesis as *exactly* true: for example, the probability of a male child will in practice deviate from 0.5 for many biological reasons. When we test such a hypothesis in the frequentist manner, we are effectively asking not whether a null hypothesis is really true, so much as whether the extent of our evidence is sufficient to detect that it is not. For such situations, a Bayesian analyst would be more likely to test a composite hypothesis like $H_0 : p \leq 0.5$ in order to detect an excess of male children. For a discussion of this approach and a comparison with the issues arising in frequentist hypothesis testing, see Armitage et al. (2002).

There are situations, however, where there is a real interest in the probability that a simple hypothesis $H_0 : \theta = \theta_0$ is exactly true, and in this case we must resort to a prior distribution which assigns positive probability at least to the point θ_0. The simplest situation is to test H_0 against a simple alternative $H_1 : \theta = \theta_1$, for which we need a discrete two-point distribution on the parameter values $\{\theta_0, \theta_1\}$. The probability that H_0 is true is then self-evidently $\mathrm{P}[\theta = \theta_0]$. Because the marginal distribution of the data cancels out from the two denominators involved, Bayes' theorem tells us that the posterior probabilities of H_0, H_1 have a ratio given by:

$$\frac{\mathrm{P}[H_0|x]}{\mathrm{P}[H_1|x]} = \frac{\mathrm{P}[H_0]}{\mathrm{P}[H_1]} \times \frac{L(\theta_0; x)}{L(\theta_1; x)} = \frac{\mathrm{P}[H_0]}{\mathrm{P}[H_1]} \times R(\theta_0, \theta_1; x), \text{ say.} \tag{26.35}$$

Each of the ratios of probabilities gives us the odds in favor of H_0 and (26.35) shows that the effect of observing the data x is to multiply the prior odds in favor of H_0 by the likelihood ratio $R(\theta_0, \theta_1; x)$, which is also known as the *Bayes factor* in this context.

This can be a useful way of weighing evidence presented in terms of a hypothesis test but appears at first sight to apply only to a single simple alternative H_1. However, the same result holds for *any* alternative value θ_1 of θ so that we can calculate the posterior odds as a function of the prior odds and θ_1. For each θ_1, we have $\mathrm{P}[H_1] = 1 - \mathrm{P}[H_0]$ and hence

$$\mathrm{P}[H_0|x] = \frac{\mathrm{P}[H_0] \times R(\theta_0, \theta_1; x)}{1 + \mathrm{P}[H_0] \times (R(\theta_0, \theta_1; x) - 1)}. \tag{26.36}$$

Example 26.20. Probability of raised risk near a nuclear installation
There is much concern about the possibility that there is an excess risk of leukemia in young people near nuclear installations in the UK and elsewhere. The issue was first raised as a result of a television program in 1983 concerning cancer in the village of Seascale in Cumbria, near the Sellafield nuclear plant. Numerous investigations have taken place since, for example, a study by Draper et al. (1993), which reported five cases of leukemia and non-Hodgkin lymphoma under the age of 25 in Seascale during the years 1963–1983. Using national cancer rates, the expected number of cases was calculated to be $\lambda_0 = 0.493$, so it would be reasonable to use the Poisson distribution to test $H_0 : \lambda = \lambda_0$ against an alternative $H_1 : \lambda = \lambda_1 = \mathrm{RR}\times\, \lambda_0$, where RR is the "relative risk" or ratio of the risk in Seascale to the national average. This gives a one-sided p-value of 0.000161, which should be regarded as highly significant and leading to rejection of H_0, though we should remember that this is an observation selected from a large number of possible combinations of time, place, and diagnosis. In any case, we reemphasize that the p-value does not tell us the probability that H_0 is true; this intrinsically Bayesian question can be arrived at via (26.36) as follows. From (26.3), we know that the ratio of the likelihoods under H_0 and H_1 is given by

$$R(\lambda_0, \lambda_1; x) = \exp((\mathrm{RR} - 1)\lambda_0) \times \mathrm{RR}^{-x}. \tag{26.37}$$

Substituting this into (26.36) allows us to calculate $\mathrm{P}[H_0|x]$ as a function of RR and $\mathrm{P}[H_0]$, and contours of this function are depicted in Fig. 26.16. Suppose, for example, that we wish to test H_0 against the hypothesis that the true relative risk is $\mathrm{RR} = 2$. In the absence of a belief in one hypothesis rather than the other, we might take $\mathrm{P}[H_0] = 0.5$. From the contour plot – or directly from (26.36) – we obtain a posterior probability of $P[H_0|x = 5] = 0.049$. The observed count of $x = 5$ is of course substantially greater than λ_0 and the posterior probability of H_0 is reduced accordingly.

26.7.6 Other Bayesian Methods

The Bayesian paradigm has proved very fruitful, and there are numerous variations and extensions of the themes we have illustrated in this section. We can only introduce here a few possibilities that the epidemiologist is likely to meet.

There are circumstances in which frequency data are available that provide information on prior distributions. An important medical example arises in the area of diagnosis, where H denotes the hypothesis that a patient has a particular disease and D represents diagnostic information; prior probabilities can then be objectively estimated from disease frequencies in the population. This approach is known by the term *Empirical Bayes* (EB). The argument can be given a frequency interpretation, and for this reason, some authors regard EB methods as not really Bayesian at all, restricting the scope of the latter to problems where priors are inherently subjective. A classical example of an epidemiological EB analysis is described in chapter ▸Geographical Epidemiology of this handbook.

Hierarchical modeling is also a feature of Bayesian analyses. Instead of choosing the parameters of the prior distribution subjectively, they are themselves given suitable prior distributions: the parameters of the latter are known as *hyperparameters*. The issues that arise in the first place are transferred to the second level of prior distribution; in principle, the process can be extended, but it is not clear that this is a useful operation. As we have argued above, there is no such thing as an

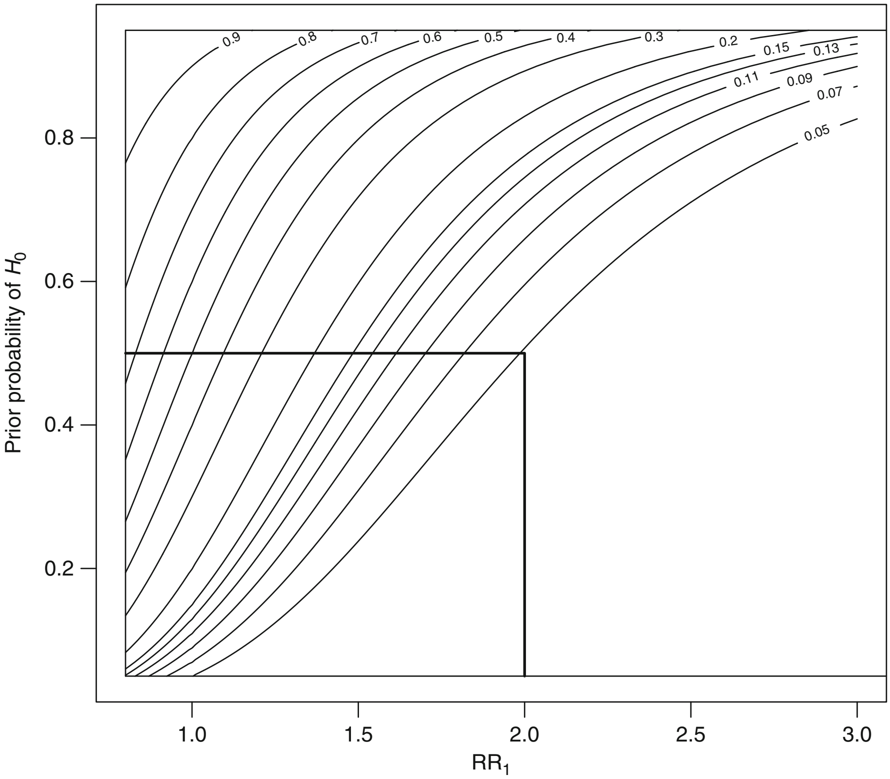


Fig. 26.16 Contours for the posterior probability of $H_0 : \lambda = 0.493$ in the Seascale data, given as a function of RR_1, the alternative value of the relative risk, and the prior probability of H_0

information-free assumption about prior distributions, and the information that is indirectly conveyed by complex Bayesian assumptions is not generally easy to assess: it is certainly not safe to assume that it can be ignored.

Highly complex models of this kind clearly pose computational challenges, and the Bayesian analyst frequently has recourse to simulation or "Monte Carlo" methods. In particular, a class of methods known as Markov chain Monte Carlo is available for estimating posterior distributions using sophisticated ideas that are beyond the scope of this chapter. The interested reader will find a comprehensive introduction in Gamerman and Lopes (2006). The methods have become very popular because of their apparent ability to model problems of arbitrary complexity and to incorporate any unknown quantities, such as hyperparameters and missing data. There is a computing system known as BUGS (Bayesian inference Using Gibbs Sampling) available on the website of the MRC Biostatistics Unit at Cambridge on http://www.mrc-bsu.cam.ac.uk/bugs/, which implements these models in a way that removes the need for any theoretical knowledge or understanding. The reader is warned, however, that this very accessibility represents a real danger for the unsophisticated user, as does the difficulty of interpreting the results.

The Bayesian paradigm is undoubtedly useful, however, for the insight it gives into the nature of statistical analysis and for enabling us to understand such problems as *model uncertainty* – that is, the extent to which we are injecting extra information into our data analysis by selecting a particular model – and the effect of making different model choices.

26.8 Further Reading

Throughout this chapter, there have been numerous references to the book by Armitage et al. (2002), which provides a wide-ranging but non-mathematical introduction to statistical methods, and also to the underlying probability theory and the general concepts of statistical inference. Many of the ideas apply across the field of medical science, of course, and the chapter on epidemiology in Armitage et al. (2002) employs a much wider definition of the subject than we have taken for this chapter. A more mathematically ambitious but clear account of statistical inference is given in Rice (2007), though there is less emphasis on methods than in Armitage et al. (2002). Clayton and Hills (1993) give an account of models in epidemiology from a likelihood standpoint. Armitage et al. (2002) have two useful chapters addressing Bayesian ideas, while Lee (2004) presents the theory at a mathematically reasonably level. At a substantially more advanced level, Cox (2006) discusses more difficult problems in the theory of statistical inference in a book that integrates classical and Bayesian ideas.

26.9 Conclusions

We have identified three different though connected approaches to statistical inference: the frequentist approach provides test procedures and estimates of parameters, but the calculation of probabilities is restricted to the description of the long-run performance of the procedures used; the likelihood approach allows a meaningful measure of how likely parameter values may be in terms of what we have called a plausibility fraction; only the Bayesian approach is capable of relating this quantity to probability statements about hypotheses or parameters.

On the face of it, the first and third of these at least are attempting fundamentally different epistemological operations. Frequentist methods may be regarded as strictly *analytical*: the estimation of parameters in a model for the data is merely a mathematical codification of the relationships that we think exist between variables in our data, while the construction of confidence intervals and the execution of significance tests are merely methodologies for assessing the role of chance in this process. Bayesian methods, on the other hand, are essentially *synthetic*: they are integrating the evidence from the data with our knowledge of the underlying scientific theory. Both operations are sensible parts of the process of scientific inference and it is arguably a mistake to put them into any kind of antagonism or competition. Doing so invites the very confusion of ideas that bedevils statistical inference: the frequentist who believes that her data represent a universal truth about her field of science is as much in error as the Bayesian who believes he is achieving a true analysis, with all the overtones of precision and objectivity that the word implies.

In one sense, the Bayesian is a statistician who believes that the integration of evidence of essentially different kinds can usefully be quantified: the extent to which this is really true will surely depend on the context. The use of subjective probability, for example, through the nomination of prior limits or odds, is dependent on an essentially frequentist view of probability. It is an interesting conjecture as to whether a "naive" mind, with no experience of any kind of repetition of experience, could usefully attach any meaning to a subjective probability merely as a result of an understanding of the axioms of probability. This difficulty parallels that of assigning any particular significance to a given plausibility or likelihood ratio in the absence of any way of relating them to a frequency theory.

That is not to say that the Bayesian approach is of no value but rather to say that it is not the fundamental solution to the problem of scientific inference that it is often held to be. This author's own view is that its value lies more in the enlightenment we gain from the Bayesian paradigm than from its use in data analysis. No doubt, opinions will continue to differ on this question, but it is to be hoped that the antagonism of the ideas to one another will become less pronounced since it is so often counterproductive.

References

Agresti A, Gottard (2007) A Nonconservative exact small-sample inference for discrete data. Comput Stat Data Anal 51:6447–6458

Armitage P, Berry G, Matthews JNS (2002) Statistical methods in medical research, 4th edn. Blackwell, Oxford

Bortolussi R, Krishnan C, Armstrong D, Tovichayathamrong P (1978) Prognosis for survival in neonatal meningitis: clinical and pathologic review of 52 cases. Can Med Assoc J 118: 165–168

Clayton D, Hills M (1993) Statistical models in epidemiology. Oxford University Press, Oxford

Cox DR (2006) Principles of statistical inference. Cambridge University Press, Cambridge

Czerniawska J, Bielen P, Plywaczewski R, Czystowska M, Korzybski D, Sliwinski P, Gorecka, D (2008) Metabolic abnormalities in obstructive sleep apnea patients. Pneumonol Alergol Pol 76(5):340–347

Draper GJ, Stiller CA, Cartwright RA, Craft AW, Vincent TJ (1993) Camncer in Cumbira and in the vicinity of the Sellafield nuclear installation, 1963–90. Br Med J 306:89–94

Everitt BS (1994) A handbook of statistical analyses using S-PLUS. Chapman and Hall, London

Gamerman D, Lopes HF (2006) Markov chain Monte Carlo: stochastic simulation for Bayesian inference, 2nd edn. Chapman and Hall/CRC, London

Greenland S (2006) Bayesian perspectives for epidemiological research: I. Foundations and basic methods. Int J Epidemiol 35:765–775

Hoenig JM, Heisey DM (2001) The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat 55(1):19–24

Lee PM (2004) Bayesian statistics: an introduction, 2nd edn. Arnold and Oxford University Press, New York

Marriott FHC (1979) Barnard's Monte Carlo tests: how many simulations? Appl Stat 28(1):75–77

Myint PK, Sinha S, Wareham NJ, Bingham SA, Luben RN, Welch AA, Khaw KT (2007) Glycated hemoglobin and risk of stroke in people without known diabetes in the European Prospective Investigation into Cancer (EPIC)-Norfolk prospective population study – A threshold relationship? Stroke 38(2):271–275

Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond A 231:289–337

Rice JA (2007) Mathematical statistics and data analysis. Third edition. Thomson/Brooks Cole. Belmont, California

Simon D, Senan C, Garnier P, Saint-Paul M, Papoz L (1989) Epidemiological features of glycated haemoglobin Alc-distribution in a healthy population: the Telecom study. Diabetologia 32: 864–869

Washburn TC, Medearis DN, Childs B (1965) Sex differences in susceptibility to infections. Pediatrics 35:57–64

Wilcoxon, F (1945) Individual comparisons by ranking methods. Biom Bull 1:80–83

Data Management in Epidemiology

27

Hermann Pohlabeln, Achim Reineke, and Walter Schill

Contents

H. Pohlabeln (✉) • A. Reineke • W. Schill
Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiolgy - BIPS, Bremen, Germany

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_48,

27.1 Introduction

Data in epidemiological studies are obtained from several sources such as questionnaires, medical records, medical devices, or laboratory tests. Data management includes the transfer of such data into a (central) database as well as all subsequent processing activities and quality control. This chapter describes the most essential steps related to the collection, entry, storage, transport, cleaning, maintenance, and statistical processing of epidemiological data, which embrace all steps from raw data to the final dataset for statistical analysis.

The main objective of an effective data management is the accurate conversion of study subjects' information into an error-free database, which is easy to handle for all subsequent analyses. Such a database builds the foundation for all statistical analyses and the conclusions drawn from the study. It is quite evident that such inferences are just as good as the quality of data. Data of poor quality, containing a great amount of random noise, decrease the power of a study and can cause type II errors (Whitney et al. 1998).

Several papers describe quality assurance and quality control procedures of clinical trials with focus on the necessity of study protocols and training of study personnel (Gassman et al. 1995; Prud'homme et al. 1989), but there is less literature in the context of epidemiological studies, where most of the corresponding data are based on the collection of personal information from individuals. Such data may be obtained via questionnaires, physical examinations, clinical laboratories (biological/genetic markers), or readouts of (medical) measurement devices. Furthermore, routine data that are mainly collected for administrative purposes are an issue of growing interest in epidemiology. Typical sources are vital and medical records, that is, disease registries, vital statistics, notification systems (especially for infectious diseases), clinical records (e.g., hospital discharge diagnoses), and health insurance claims data. Also aggregate level data may be used like census data or Bureau of Labor Statistics (BLS) data as well as information provided by geographic information systems (GIS).

Although it is desirable to avoid any errors, they occur even with careful data collection and entry. Therefore, given the huge amount of work, money, and time usually involved in the collection of data for epidemiological studies, it is crucial to avoid additional errors during the transfer of data from paper-based forms or measurement devices to digital data. Even if many data errors will be detected by chance at different stages of data management and analysis (other than data cleaning), it is more efficient to screen systematically for errors before the statistical analysis is carried out (Van den Broeck et al. 2005).

A wide range of software products is available for data entry. Most software packages support the user during the process of data entry. For numerical variables, the data entry can be restricted to plausible ranges only and for alphanumerical variables to valid codes. It is also possible to force data entry for compulsory fields. Regardless of any procedure to assure data quality during entry, corresponding errors can be minimized by entering data twice by different persons and by subsequent comparison of both entries. Discordant values are then cross-checked against the original document, and the valid value is kept in the corrected dataset.

However, even with the most careful verification of data entry, it is necessary to perform additional checks before beginning data analysis. Plausibility checks can be run automatically and comprise queries regarding the plausible range as well as cross-checks of variable values. Finally, sometimes "…a data review meeting before database lock, which involves all parties, to discuss the data and remaining implausibilities (with the consequence of a possible exclusion of data) is also recommended" (Theobald et al. 2009). By all means, all plausibility checks are to be determined before any analysis of data starts.

This chapter is organized in four main sections, based on our experience in data management and on several publications in which many of the concerns in this field have been described in detail. We introduce the issue on how to set up a data management infrastructure to handle data from diverse sources. The second section addresses the entry, storage, and transfer of data, followed by a section focusing on the main steps of data cleaning, that is, to detect, correct, or discard invalid values and records from a database. The last section deals with key concerns on how to prepare the final analysis dataset containing all variables, original, and derived, on which all calculations are performed.

27.2 Setting Up the Data Management Infrastructure

Epidemiological studies vary considerably with respect to use of data sources, recruitment of participants, and study purpose. Data may be collected directly from study subjects or may be retrieved from secondary data such as administrative databases, clinical records, or registries. Such a collection of data as well as the data management procedures should be planned carefully and timely to avoid waste of time and invalid data. Retroactive correction of mistakes requires additional time and manpower and is often impossible. The complete set of software tools and devices should be well tested, and their handling should be trained. All procedures have to be laid down, for example, as standard operation procedures. Sufficient time and resources to plan and set up all data management procedures should be allocated from the very beginning of a study.

This section describes the necessary infrastructure and data management procedures to handle data from such diverse sources.

27.2.1 Requirements, Tools, and Set-Up

Data sources and used methods of data collection or retrieval have to be defined before planning the data management (see chapter ▸Epidemiological Field Work in Population-Based Studies of this handbook). Procedures differ depending on the source of the data, the data collection procedures, and the measurement devices used. The choice of tools and software may be influenced by the skills, knowledge, and experience of the staff.

27.2.1.1 Source of Data

Secondary Data When data are extracted from an existing database, data management is comparatively simple. The structure of the data is predefined; they are grouped in tables, the relations between these tables are defined and the types of the variables (e.g., numeric or character) are given. Problems can arise if the data have to be transferred to a different database system. Pitfalls might be the different representation of variables in different systems (e.g., length of an integer value) or the coding of missing values.

If the study data arise from different sources with differing data structures or if the given structure cannot be used, the data management is more demanding. In this case, data have to be selected, linked, transformed, and integrated into one uniform structure. In order to combine two datasets, key variables are required to merge corresponding data records. For example, usually it is not feasible to merge records from a cancer registry and a clinical dataset only by the name of a patient. Since different patients may have the same name, further information is required (e.g., date of birth) in both datasets to generate a unique identifier to merge corresponding records accurately. If a unique identifier is not available, it will be complicated, if not impossible, to combine the data.

Another challenge in combining existing datasets is the use of differing classifications for coding. For example, diseases are usually classified according to the International Classification of Diseases (ICD; World Health Organization 2009). However, the ICD revision used for coding may differ between data sources, for example, ICD Version 9 or 10. Another simple example is the use of different codes for the same variable values (e.g., male/female coded as 0/1, 1/2, or m/f). Such differences have to be identified, and the coding has to be harmonized.

Primary Data If data are newly collected by interviews or measurements, the data management requirements have to be formulated while planning and designing the study. Selection and configuration of instruments and measurement devices should be done bearing the final data integration in mind. A unique identifier (see Sect. 27.2.3.2) on all forms, questionnaires, and measurement outputs allows to merge all records of a given study participant. Coding of character variables should be done according to predefined classifications. It is advisable that all requirements are known and that the used software tools, coding systems, and structure of the resulting database are defined before the data collection starts. A change in a questionnaire after commencement of data collection usually leads to immense difficulties in the data management.

27.2.1.2 Methods for Data Collection

A wide variety of techniques for data collection is in use: paper questionnaires or touchscreens for self-completion or for use by an interviewer in face-to-face interviews, computer-assisted telephone (CATI), or computer-assisted personal (CAPI) interviews, and web-based solutions where the study participant enters the data directly in a web-form hosted on the server in the study center. All of them have strengths and limitations. The choice depends on the requirements of the study, the technical equipment, the budget, the skills of the personal, and the characteristics of study participants. For example, it will be a challenge to motivate elderly persons to enter data in a web-form. However, entering data directly into a central database eliminates problems from merging and updating isolated databases. The final choice of methods and procedures should therefore be discussed in the study team, where the person responsible for data management can contribute possible solutions and ensures consideration of IT aspects.

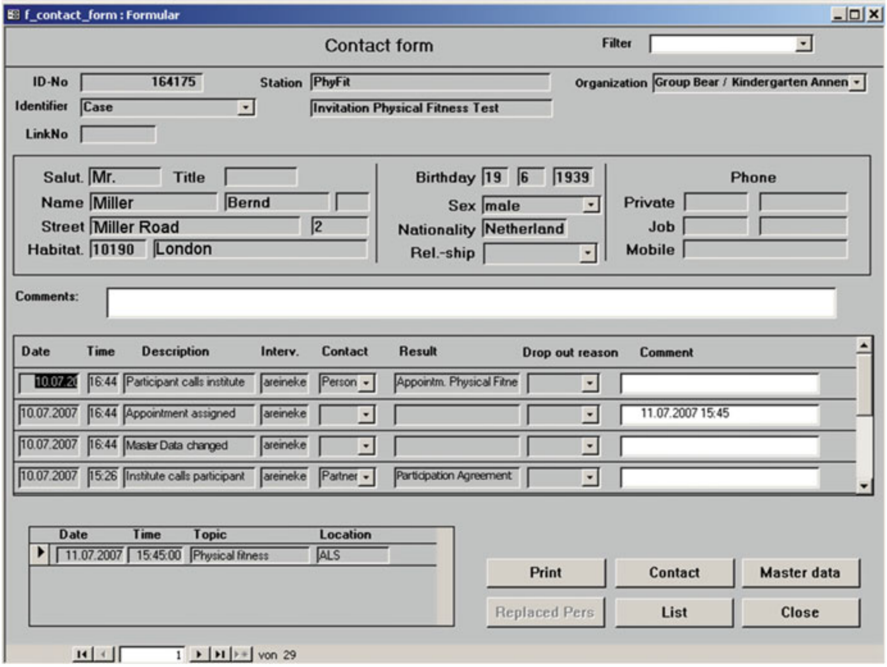
27.2.1.3 Measurement Devices

When using measurement devices like sphygmomanometers, accelerometers, or ergometers, additional data management issues have to be considered. For example, the possibilities of downloading and processing of the collected data must be verified. Will it be possible to process the data with standard programs, or is there a need for specialized software? In any case, the configuration of the devices must be checked and tested. In most cases, the interval of measuring must be defined, or the collection of additional parameters must be switched on or off. It is absolutely necessary to define a standard configuration and a robust procedure to ensure an identical set-up if several devices are used and to enable a rapid replacement of lost or broken devices.

27.2.1.4 Software Tools to Support Fieldwork and Data Management

When collecting data in a large epidemiological study, specific IT tools can help to recruit study participants, to clean the data, or to manage biological material.

Tools to Manage the Recruitment of the Study Participants Surveys are an integral part of most epidemiological studies. The overall quality of the study does not only depend on the appropriateness of the study design and the validity of the measurements, it also depends on the fieldwork. Usually the complex fieldwork procedures are not well documented. Information about the recruitment of potential study participants, their tracking, and the characteristics of non-responders are extremely important to assess potential selection bias and to judge the generalizability of the study results (Morton et al. 2006; Olson et al. 2002; Tooth et al. 2005; Vrijheid et al. 2009). Besides the response proportion there are many other aspects of the fieldwork that should be documented to allow assessment of the study quality (Hartge 2006). Mode, date, time, and outcome of each contact with potential study participants must be documented to monitor adherence to defined contact procedures and to analyze reasons for non-response. A field management tool will reduce the required resources for management and documentation drastically. One approach is to set up a small database with the contact details of all potential study participants to control contact mailings and reminders. Appointments for interview and examination could be documented in a separate system, for example, in a timetable. A more sophisticated approach uses a specialized system integrating all required functions like MODYS – a modular control and documentation system for



ID-No 164175
Station PhyFit
Organization Group Bear / Kindergarten Annen
Identifier Case
Invitation Physical Fitness Test
LinkNo

Salut. Mr.
Title
Name Miller Bernd
Street Miller Road 2
Habitat. 10190 London
Birthday 19 6 1939
Sex male
Nationality Netherland
Rel.-ship

Phone
Private
Job
Mobile

Comments:

Date	Time	Description	Interv.	Contact	Result	Drop out reason	Comment
10.07.20	16:44	Participant calls institute	areineke	Person	Appointm. Physical Fitne		
10.07.2007	16:44	Appointment assigned	areineke				11.07.2007 15:45
10.07.2007	16:44	Master Data changed	areineke				
10.07.2007	15:26	Institute calls participant	areineke	Partner	Participation Agreement		

Date	Time	Topic	Location
11.07.2007	15:45:00	Physical fitness	ALS

Fig. 27.1 MODYS main contact form. *Top*: basic information of the participant, such as ID number, name, and address. *Middle*: log of all performed tasks and contacts. *Bottom*: appointments and buttons to document additional contacts

recruitment of study participants in epidemiological studies (Reineke et al. 2014). MODYS was designed to standardize and record all participant contacts during the recruitment phase of a study. It provides functions to store the contact data for each potential study participant, to generate mailings and reminders for individuals or groups, to manage appointments (e.g., for interviews and examinations), and to document each contact with date, time, involved persons (staff member, participant, see Fig. 27.1), and outcome. MODYS can be configured according to the specific contact procedures in a study. The course of the recruitment is arranged in the system in a tree-like structure reflecting the stage of recruitment for each potential study participant. After all activities foreseen for a given stage are completed (e.g., a letter was printed and the corresponding documentation was generated), the potential participant enters the next stage in the tree. The system thus forces that the recruitment follow a predefined process in a controlled way and ensures proper documentation of each step. Thus, MODYS allows to monitor the survey progress and to optimize the workflow.

Tools for Data Cleaning Data cleaning is particularly demanding, especially in multicenter studies. To find implausible or wrong values, queries have to be generated by standardized plausibility checks of raw data, and resulting corrections must be carried out and documented. Typically, lists of queries will be generated in the data coordinating center (DCC), sent to the study centers, and processed, and the results will be sent back to the DCC. Development of a correction database including all queries and questionable values can facilitate the necessary procedures. In each study center, all corrections have to be entered in this database. Finally the corrected data will be transferred to the DCC where they are easily integrated in the central database. Documentation of all changes is included in the correction database. Section 27.4 gives an overview of procedures for checking and cleaning of data.

Tools to Manage Biological Material Collection, processing, shipment, storage, maintenance, and supply of biological material should be documented for each single sample. Biobanks are used to manage collections of biological materials such as blood, urine, and cell cultures and to record storage locations throughout the workflow including volume tracking. Procedures and conditions (e.g., dates of freezing and thawing, storage temperatures) should be controlled and recorded throughout the storage period to safeguard storage conditions and to allow evaluation of negative effects of unforeseen deviations from standard operating procedures.

A broad range of systems is available from small spreadsheet-based solutions up to databases to store information about thousands of biosamples with an integrated retrieval function and a web-based user front end. For each single sample, a tool to manage biological material should provide information about:

- Storage location (tube, container, freezer, room, building…)
- Type and volume
- Date and time of freezing/collection
- Freezing temperature and all deviations
- Date and time of sample shipments
- Date and time of thawing

27.2.2 Data Protection

Working with personal information given by study participants is a privilege that must be respected. Over and above legal requirements, epidemiologists should seek to minimize (or better exclude) the risk of disclosure of personal information through appropriate technical solutions and procedures.

Procedures and techniques to ensure the confidentiality of the study data depend on the type of the study, source of data, and local laws and specific regulations applying to these data. Although always the same general principles of data protection have to be considered, a universal data protection procedure, suitable for any kind of study, does not exist. Each study needs its own adaptation. Some general rules and guidelines have been worked out for the IEA-Guidelines for Good Epidemiological Practice (IEA International Epidemiological Association 2007).

Central principles are described in the International Ethical Guidelines for Epidemiological Studies (CIOMS 2008).

27.2.2.1 Personal Identifiers

Personal identifiers such as names, addresses, phone numbers, social security numbers, or email addresses must be detached from other data of a given study subject as soon as possible after data collection. If their storage is necessary, for example, for follow-up, personal identifiers have to be kept separately from study data under lock and key. Data containing personal identifiers may only be used if it is absolutely necessary where access to these data requires special authorization of personnel that is legally committed to observe all confidentiality rules. In general, such data are needed during the data collection, for example, to contact study participants as part of the data cleaning process or in longitudinal studies to follow-up individuals. Separation of study data and personal identifiers needs to be reflected in the organizational structure, where only field staff may have access to identifiers while scientists are allowed to access pseudonomized study data only. A pseudonym, that is, a study-specific identification number, is used to allow linkage of epidemiological study data and personal identifiers (see Fig. 27.2). Sometimes it can be necessary to use several different pseudonyms for one and the same person. If, for example, DNA is collected, it is now common practice to provide these data with another pseudonym than the study data. This is done to prevent misuse of DNA or re-identification of study subjects by their genetic profile (see Sect. 27.2.3.4).

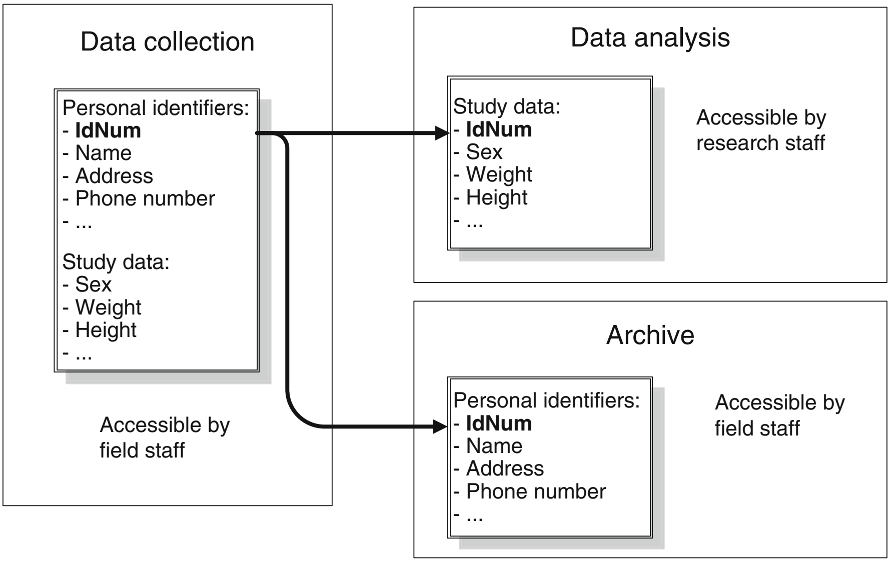


Fig. 27.2 Scheme illustrating of the separation of personal identifiers and study data

27.2.2.2 Data Protection on Mobile Devices

Data containing personal identifiers should not be stored on computers outside the study center (e.g., a laptop used at home). In case the data have to be used (or collected) by mobile storage media or if data have to be transferred, extra safeguarding procedures are necessary. Confidential data should never be sent in any electronic form (by mail or as CD) unless they are encrypted. Additional procedures include shipment by personal courier or by special delivery mail. If personal data have to be used on laptops in the field, they have to be protected on the hard disk. Several software products and, recently, more hardware-based solutions are available to encrypt data on a hard disk automatically. External hard disks with an integrated encryption are inexpensive and can thus be used in each study. If such a disk gets lost, the data will be secure: without a password or key, the data on the disk cannot be read. Another possibility is to use encrypted partitions on a laptop. User-friendly products like BitLocker (integrated in Microsoft Windows) or TrueCrypt (2012) are widely used.

27.2.2.3 Authorization of Data Access

Access to the study data and especially to the personal identifiers has to be restricted to authorized staff members. Different access rights should be granted depending on the role of each person in the project (e.g., researcher, interviewer, data manager; cf. Fig. 27.2). Access rights have to be withdrawn if a staff member leaves the study team.

Today, every computer system provides mechanisms based on user accounts and corresponding access rights to restrict the access to files or directories. Different access rights have to be granted to each user. An individual user account is more useful than an account for a group of users. In contrast to that, the access rights for each user depend on the duties and tasks, that is, the individual role of the user. For example, since interviewers need other access rights than statisticians, two roles, one for interviewers (*R_Interv*) and one for statisticians (*R_Stat*), can be defined. User accounts of statisticians will then receive the role *R_Stat* and user accounts of interviewers the role *R_Interv*. Access can then be withdrawn by deletion of the link between a user account and the role instead of deleting several access rights.

The need for confidentiality and data protection does not end with the completion of a study because data must be kept under lock and key for several years beyond the study period to allow verification of published results. Backup copies (see Sect. 27.3.3) must be subject to the same degree of data protection.

27.2.3 Set-Up and Implementation

The set-up and implementation of the data management routines could start if all general definitions have been made and all requirements are known. The choice of procedures and tools may depend on the size and complexity of the study.

27.2.3.1 Creating a Database

Collected data are usually entered and stored in a tabular format, apart from special forms, for example, data stored in images. In the case of a questionnaire, each respondent is represented by a row in the table, a so-called record. Each record contains a series of variables, one for each item in the questionnaire. Separate tables are used to store data derived from different sources.

The methods and functionalities to store and manage diverse tables depend on the software. The decision for a software product depends on the complexity of the study and the instruments used for data collection. In a small study it is sufficient to store data in a spreadsheet-like flat file. In a large study, a database management system (DBMS) which integrates several tables may be more efficient, although its set-up requires more resources.

Management of a small amount of data in a flat file where each variable defines a column is easy. The name of the variables can be entered in the first row of the table followed by the data of the participants in subsequent rows (see Fig. 27.3). Since the allowed length of variable names is limited, they are usually abbreviated. Full names and further specifications like variable format and length have to be specified in a data dictionary (see Sect. 27.2.3.5).

27.2.3.2 Complex Data Model

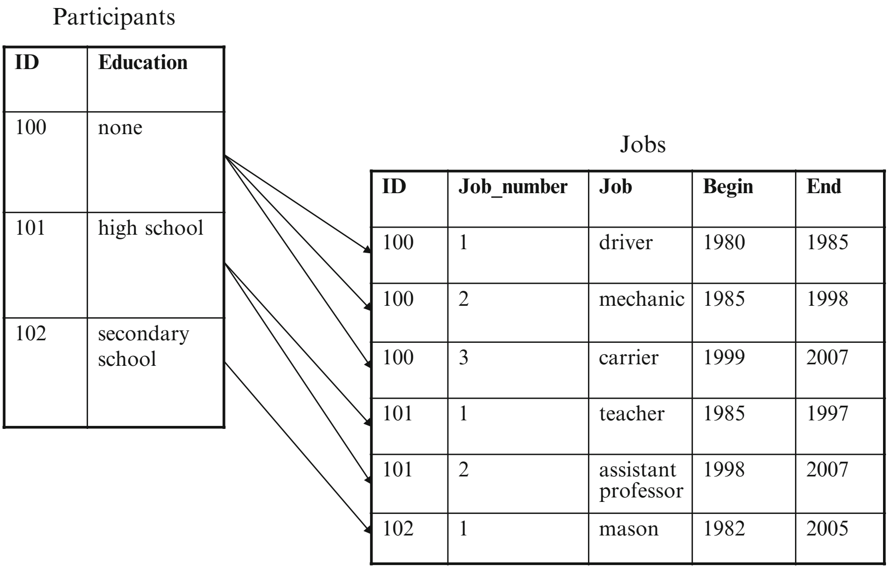
If the data model becomes more complex, however, it will be difficult to store the complete data of an individual participant in a single row. If, for example, a complete job history is recorded, it is bad practice to reserve as many columns (variables) in a spreadsheet as there are needed to record the maximum number of job periods reported in a given study. For each job period, several variables have to be recorded like job title, start and end date, branch of industry etc.; hence, a huge number of columns have to be foreseen to record such phase-related information. Not least it is difficult to estimate the maximal number of columns needed. If too few columns are foreseen in the database, it would have to be changed while the study is ongoing. If too many columns are reserved, most of them will remain empty.

Allowing a separate record for each job period is more flexible. This can be achieved by the use of a second spreadsheet with a link variable, for example, the identification number, to the first one, where the first one entails basic characteristics of the study subject like age and sex, while the second one stores each job period in

Participants and Jobs

ID	Education	Job_1	Begin_1	End_1	Job_2	Begin_2	End_2	Job_3	Begin_3	Job_3
100	none	driver	1980	1985	mechanic	1985	1998	carrier	1999	2007
101	high school	teacher	1985	1997	assistant professor	1998	2007	n.a.	n.a.	n.a.
102	secondary school	mason	1982	2005	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

Fig. 27.3 Flat file model to record level of education and job histories (job title, begin and end)



Participants

ID	Education
100	none
101	high school
102	secondary school

Jobs

ID	Job_number	Job	Begin	End
100	1	driver	1980	1985
100	2	mechanic	1985	1998
100	3	carrier	1999	2007
101	1	teacher	1985	1997
101	2	assistant professor	1998	2007
102	1	mason	1982	2005

Fig. 27.4 Database model organized in linked tables to store level of education and job histories (job title, begin, and end)

a separate row. Each study participant can get as many rows as needed to record all his/her job periods (see Fig. 27.4).

27.2.3.3 Appropriate Software

The selection of the appropriate software to conduct a study depends on the amount, type, and form of data to be collected. If data from hundreds or more persons have to be collected or if complex relationships between variables have to be mapped, a database management system is required. If the structure of data is less complex, a more simple and easy to use system should be chosen. However, the selection should not only depend on the technical possibilities and requirements. The expertise of the staff is also an important criterion. Even if a system is semi-optimal, it can be a good choice if it is well known by the staff.

Database Management Systems In the context of DBMS, the connections between spreadsheets as shown in Fig. 27.4 are called relations. A DBMS is able to connect records between tables by a link variable, for example, the ID (identification number) of the participant, and provides functions to store and retrieve the data. A DBMS thus avoids redundancies, reduces the possibility of errors, and eases data management procedures. Several DBMS products are available, starting with systems like Microsoft ACCESS®, which are easy to use but limited in their ability to manage big amounts of data. More specialized systems are ORACLE®, Microsoft SQL-Server®, IBMsDB2®, or MySQL®, designed to manage big and complex databases.

Other than flat file solutions, most DBMS can be used by several persons simultaneously, that is, different data entry clerks could enter data at the same time. This can accelerate data entry, while all plausibility checks are done against the complete database. If, instead, data were entered in parallel into independent copies of the database, some checks would not work. For example, the check whether a given questionnaire has already been entered could only be done separately for each single copy. Then, inadvertently, a single questionnaire might be entered twice resulting in a duplicate in the final dataset. In general a DBMS stores all data in one big file on the hard disk. Within the DBMS the data are organized in separate tables (spreadsheet = table). Other systems use several files to store the data (spreadsheet = file).[1]

Statistical Systems To enter, store, and manage the data directly in the system used for analysis is another widely used solution. Systems like SAS®, SPSS®, STATA®, or R® provide the capability to enter and manage the study data. The data can be analyzed directly without any transformation, but none of these statistical packages provide as many functionalities for data management as a DBMS.

Systems Used in Epidemiological Studies Some specialized systems combine several functions needed to conduct an epidemiological study. Such systems provide, for example, a database engine to store the data, a form generator to build data entry forms, and a component to analyze the data. One of those systems is Epi Info™ developed by the CDC Centers for Disease Control and Prevention (2011). Epi Info™ is a public domain software package designed to provide a base framework for conducting an epidemiological study. Some other systems with differing functionalities can be found in the internet, for example, EpiData (EpiData Association 2010), or OpenEpi (Dean et al. 2011). Depending on the requirements of the planned study, these systems are certainly candidates to consider.

27.2.3.4 Configuration

The number of tables required to store and manage the data and their relations has to be defined. Each item of a questionnaire has to be assigned to a variable. The number of variables and their types (e.g., numeric or text) directly depend on the questionnaire, but the method to store it in the computer and the conventions to name them are predefined by the data management system. For example, the possibilities to name a variable can be limited in length or by the disposable characters.

Unique Identifiers Unique identifiers (UIDs or IDs) are used to identify participants, to link separate records and biological samples of the same participant, and, more technically, to define the relations between tables in the database. In Fig. 27.4,

[1] The definition is simplified. From the perspective of a database administrator, for example, an ORACLE database consists of several files, and it is also possible to store more than one spreadsheet in one R file.

for example, the variable *ID* defines which rows in the table *jobs* are connected to which row in the table *participants*.

Sometimes one ID number can be used for all purposes: to link all datasets, all biological samples, and all questionnaires of an individual. In other cases, different identifiers are preferable. It is, for example, good practice to enforce data protection by using different IDs for participant data and their biological material. For example, DNA is often considered to be particularly vulnerable to misuse. Assignment of different pseudonyms for reported data and DNA samples provides an additional barrier against re-identification of study subjects by unauthorized persons.

The construction of IDs depends on the intended purpose of the study and the legal requirements that have to be observed. Natural identifiers like the name of the participant or the social security number should not to be used as an ID. As discussed in Sect. 27.2.2.1, all personal identifiers have to be separated from the other study data. Therefore, the ID that is used in the study dataset should be a pseudonym (see Sect. 27.2.2.1) to hinder direct identification of study subjects. An ID should include all information necessary but as few information as possible. A simple method to generate such an ID is to number participants consecutively as they are entered into the database. The link between the participant name and number can be stored under lock and key in a translation list or file. Without such a list, the link of a record in the database to the corresponding participant is impossible. To avoid an association between the ID and the time when the participant was recruited, a random number is useful. Such a simple ID consisting of a five-digit random number may look like the following:

ID:	1	4	5	4	6
	random number				

More sophisticated approaches incorporate additional information in the ID, like the number of a study center in a multicenter study, the number of examinations in longitudinal studies, or a characteristic of the participant (male/female, age group, company number, etc.).

To avoid errors while entering the ID into the database, a checksum should be used. In this case, the last digit of the ID is a checksum calculated out of all other digits according to a predefined algorithm (e.g., Gumm 1986). A more complex ID containing a study center number, a five-digit random number, and a checksum may look like the following:

ID:	1	2	1	4	5	4	6	8
	center		random number					checksum

In study center 12, the participant 14546 was recruited. Out of 1214546, a checksum of 8 was generated. This method to create an ID is used in the German

National Cohort Study (Wichmann et al. 2012). In general it is advisable to avoid leading zeros (e.g., in this case a study center 09). If the ID is stored as numeric variable a leading zero will be truncated.

The inclusion of characteristics of the participant into the ID should be decided with care. If, for example, the number of a company is included and the participant changes his/her employer, this part of the ID is no longer valid. To change the ID in an ongoing study is complicated and error prone. Also it is not advisable to include information that is related to exposure or outcome, like the case-control status of the participant, since this could unblind study staff and thus hinder blind assignment of codes to study subjects. An incorrect ID on a questionnaire or a sample tube causes an incorrect (or missing) link to the participant. Using a barcode scanner to read IDs automatically will reduce typing errors.

Naming Variables Each variable name identifies a unique variable in the database. Each name should

- Recollect the meaning and/or position of the question or data item in the original document.
- Be consistent and short (shorter names are easier to use, some software programs only allow eight characters, and variable names will usually be displayed in graphs).
- Use lower-case letters to prevent errors if case-sensitive software is used (e.g., UNIX systems vs. MS Windows).
- Avoid special (national) characters like a German ä or a French è.
- Not contain spaces or punctuation marks.

Variable Types Common types of variables are alphanumeric (char, varchar, text, string, etc.), numeric (byte, integer, long integer), logical, and date. The format of a variable defines the possible number of characters or digits and how it is presented by the system. A date variable is stored in the computer as an integer number (e.g., days since December 30th, 1899 + seconds since midnight) and could be presented as 10.01.2011 or as January 10th 2011 or 10 Jan 2011. Text variables may contain letters or other special characters.

A string variable of length three contains up to three characters. A numeric variable of type byte contains numbers between 0 and 254. For each variable, valid values can be defined to allow for plausibility checks which can be programmed with the data entry forms to integrate related checks. Often it is advisable to use numeric variables or to convert non-numeric information to numeric variables because they are generally more comfortable during data analysis. But sometimes data handling is easier if a variable is coded as a string variable even though it only consists of numbers (e.g., zip codes with leading zeros).

It may be suitable to use more than one variable for one data item. For example, the name of a participant is commonly defined as two variables, one for the surname and one for the family name, even if only a single text field is provided. Separation of a calendar date into three separate (numeric) variables allows entering of incomplete information. For example, if a participant only recalls the year and the month of starting a job period, the exact day may be omitted.

Table 27.1 Examples of codes for missing values

Code	Type of missing	Description
−6	Not applicable	Question not applicable for the respondent; e.g. number of deliveries of a male respondent
−7	Refused to answer	Respondent refused to answer that question
−8	Don’t know	Respondent does not know the answer
−9	Missing	Left empty although question is applicable

Coding of Variables An important issue, following the data collection, is the coding of variables, that is, the translation of questionnaire information into numerical values, representing the answers to the questions. Nominal variables allow for only qualitative classification; that means it is not possible to quantify or even rank the categories. A typical example is a question concerning the marital status. A possibility of coding would be married = 1, living in a marital-type relationship = 2, single = 3, divorced/separated = 4, and widowed = 5, ideally supplemented by special codes for missing values (see Table 27.1). The attribute ordinal refers to nominal variables with a natural order (e.g., school grades, income levels). In surveys, ordinal variables are often used to measure the respondents’ attitudes or opinions by asking the extent to which they agree or disagree to a series of statements (Likert scale 1932). A typical Likert scale would be “strongly agree (1),” “agree (2),” “not sure/undecided (3),” “disagree (4),” and “strongly disagree (5).”

Special codes are needed to document and classify non-responses, which may occur for several reasons. Knowledge on these reasons is helpful for in-depth analyses of non-response. It is therefore a good idea to distinguish between causes like “refused to answer,” “don’t know,” or “not applicable” and to assign numbers to these categories which should be outside the range of codes for valid answers (see Table 27.1).

Analysis of answers to open-ended questions is generally less straightforward since responses given as free text are not directly accessible to data analysis. In the phase of programming, it is not feasible to assign a code to all possible answers. Open-ended questions therefore need to be coded after data collection with the help of available classification systems, for example, the International Standard Classification of Occupations (ISCO) to code job titles (International Labour Office 1968), the International Standard Industrial Classification (ISIC) of All Economic Activities (United Nations Publications 1971) to the branch of industry, or, for the coding of diseases, the International Classification of Diseases (World Health Organization 2009). However, sometimes open-ended questions require the development of a separate code frame in order to assign codes to the returned plain text answers.

27.2.3.5 Data Dictionary/Codebook

All variable definitions should be documented in a list called codebook or data dictionary (see Table 27.2). The data dictionary is the blueprint to set up the database and the key document for the data management that will be used throughout the whole study (van Es 1996) for defining the data collection forms, the data entry screens, the study database, and the analysis datasets. It is one of the main sources provided to the data users to understand the meaning of the collected data. An excerpt from a data dictionary is given in Table 27.2 where the column

- *Q-No.* indicates the number of the question in the questionnaire.
- *Name of variable* shows the name of the variable in the database and should be retained throughout all steps of data analysis.
- *Description of variable* includes a short description of the variable.
- *Type and format* describes the type and the length of data (numeric, character).
- *Value list/definition* describes the valid values of the variable (range, missing codes, etc.) or refers to a more complex definition in the appendix of the data dictionary.
- *Plausibility checks* describe, possible plausibility checks to be considered during programming; complex checks are defined in the appendix of the data dictionary or in a separate document.
- *Data source* contains the name of the table in the database or the name of the file (see Sect. 27.2.3.3) where the variable is stored in.

The codebook describes both collected and derived variables. Collected variables represent original questionnaire items like weight, height, or job titles, whereas derived variables are calculated from one or more of these collected variables, for example, the body mass index (BMI) which is calculated as the ratio of body weight (in kg) over square of height (in meters), or the derived variable *list_A_dur* which is the duration (years) of employment in occupations and industries known to be associated with lung cancer, the so-called *list A* (Ahrens and Merletti 1998). The codebook connects both types of variables with their names, types, formats, and further information. The data dictionary should at least include the unique name, the type, and all permissible values of a given variable.

27.2.3.6 Creating Forms for Data Entry

Well-designed entry forms are a prerequisite for reliable data entry. Special data entry forms have several advantages over the direct entry into an unformatted table of a database. A clear and neatly arranged design of the data entry screen, a defined order of entry boxes, and hot key short cuts can reduce the error rate and accelerate data entry. A data entry form can resemble the layout of the original paper form or questionnaire. This supports the data entry operator to not accidentally skip entries as he/she is guided intuitively through the form field by field. The implementation of automatic skip patterns ensures that the cursor automatically moves to the next input field when the maximum number of characters was entered in the current input field. Skip-and-fill rules make sure that fields are filled automatically as a function of previous responses or that fields are skipped if they are not applicable. For example,

Table 27.2 Example of a data dictionary (the variables id_name and id_sname are stored in an extra table and should be kept separately from study data under lock and key)

Q-No.	Name of variable	Description of variable	Type and format	Value list/ definition	Plausibility checks (description or reference)	Data source (name of table/file)
1.	*idnum*	Participant ID in the study	Number 8.0 digits	[a]	Checksum	ident_tab main_tab
1.1	*id_name*	Participant name	String 30 char			ident_tab
1.2	*id_sname*	Participant surname	String 30 char			ident_tab
. . . ..						
4.1	*height*	Participant height (centimeter)	Number 3.0 digits	Missing = −9	Range: 130–215 or −9	main_tab
4.2	*weight*	Participant weight (kilogram)	Number 3.0 digits	Missing = −9	Range: 40–150 or −9	main_tab
4.3	*dat_d*	Participant day of birth	Number 2.0 digits	Missing = −9	Range: 1–31 or −9	main_tab
4.3	*dat_m*	Participant month of birth	Number 2.0 digits	Missing = −9	Range: 1–12 or −9	main_tab
4.3	*dat_y*	Participant year of birth	Number 4.0 digits	Missing = −9	Range: >1912 or −9	main_tab
4.4	*sex*	Participant sex	Number 1.0 digit	Male = 1 Female = 2 Missing = −9	Range: 1, 2, or −9	main_tab

(*continued*)

Table 27.2 (Continued)

Q-No.	Name of variable	Description of variable	Type and format	Value list/ definition	Plausibility checks (description or reference)	Data source (name of table/file)
4.5	*n_child*	Number of deliveries	Number 2.0 digits	Missing = −9	Range: 0–20 or −9 (valid only for females >14 years)	main_tab
. . . ..						
4.7	*edu_lev*	Educational level	Number 1.0 digit	Elem. school = 1 Mid. school = 2 High school = 3 Other school = 4 not applicable = −6 Refused = −7 don't know = −8 Missing = −9		main_tab
4.7	*edu_lev_t*	Other educational level	String 30 char			main_tab
4.1–4.2	*bmi*	Body mass index (kilogram per square meter)	Number 2.1 digits	Derived variable[a]	See appendix	main_tab
5.1–5.2	*smoke_dur*	Duration of smoking (years)	Number 2.0 digits	Derived variable[a]	See appendix	main_tab
6.4–6.2	*list_A_dur*	Duration in industries/occupations known to be associated with lung cancer	Number 2.0 digits	Derived variable[a]	0–60 or −9	jobs_tab

[a] The definition of these variables can be seen in the appendix of this data dictionary

if *Question 1* asks whether a study participant ever smoked cigarettes and *Question 2* asks for the average number of smoked cigarettes per day, then a response of "Never" to *Question 1* could automatically imply a code for "not applicable" into *Question 2* and a skip of the cursor to *Question 3*. If possible, the plausibility checks defined in the data dictionary should already be active during the data entry process. While developing forms for data entry, the following aspects should be considered:

- Instant plausibility checks during entry should be an integral component, but too many restrictions could lead to a higher error rate or impede data entry. Each invalid value will interrupt the process. Some flexibility to enter "invalid" values is therefore necessary. Any system should therefore foresee the option to check back the original data source and to revisit data entry at the same point later on.
- Key variables like IDs must be treated with maximum care. Double entry or use of barcodes may reduce errors.
- Wherever applicable, a list of valid values of a variable should be provided.
- ID of data entry clerk and date of entry should be recorded.
- Fields not used should be disabled or completely hidden to prevent erroneous entries. If, for example, the question for smoking is negated, a subsequent field to enter the number of smoked cigarettes should be disabled.
- Disabled fields should be distinguished from fields erroneously skipped by the data entry operator; they should be filled with a default value for "not applicable" (see Table 27.1).
- The data entry system should track version numbers of questionnaires or data forms (in case that instruments are updated during the study).

Data entry is a critical process. It introduces several sources of errors and should be standardized and conducted with a high degree of reliability; see chapter ▶Quality Control and Good Epidemiological Practice of this handbook.

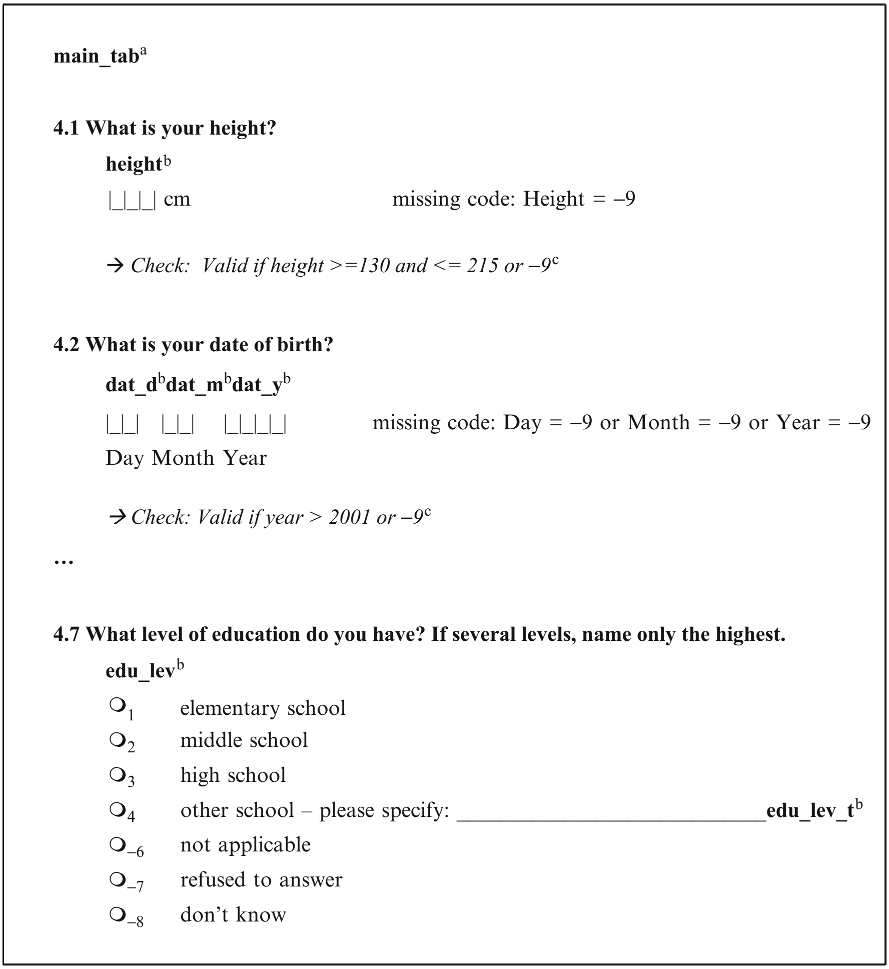
27.2.3.7 The Annotated Questionnaire

The annotated questionnaire is a paper or electronic version of the questionnaire enriched by the names of variables (used for storing the data) of questions, measurements, and other collected data items. Further annotations may relate to skip patterns, ranges of valid questionnaire responses, or interviewer instructions. It is advisable to use an electronic version of the questionnaire, for example, a file out of the word processing system or a PDF file, for annotation (see Fig. 27.5).

27.3 Entry, Storage, and Transfer of Data

27.3.1 Data Entry and Visual Editing

As described in detail in Sect. 27.2.1.2, a wide variety of techniques for data collection is in use. The entry of the data into an electronic database is the final step of data collection. Data entry depends on several preconditions and should be performed



main_tab[a]

4.1 What is your height?

height[b]

|_|_|_| cm missing code: Height = –9

→ *Check: Valid if height >=130 and <= 215 or –9*[c]

4.2 What is your date of birth?

dat_d[b]**dat_m**[b]**dat_y**[b]

|_|_| |_|_| |_|_|_|_| missing code: Day = –9 or Month = –9 or Year = –9

Day Month Year

→ *Check: Valid if year > 2001 or –9*[c]

…

4.7 What level of education do you have? If several levels, name only the highest.

edu_lev[b]

○$_{1}$ elementary school

○$_{2}$ middle school

○$_{3}$ high school

○$_{4}$ other school – please specify: ____________________________**edu_lev_t**[b]

○$_{-6}$ not applicable

○$_{-7}$ refused to answer

○$_{-8}$ don't know

[a]Name of the table or file where the data are stored

[b]Name of the variable in the table

[c]Plausibility check

Fig. 27.5 Excerpt of an annotated questionnaire

with the same care and attention as other procedures in the study. It should start as soon as possible after commencement of data collection. If data entry only starts after completion of the fieldwork, feedback of problems encountered during entry to the field team is useless, and adjustment of field procedures or amendment of instructions for the editing team (cf. visual editing) is no longer possible.

Data entry systems provide suitable data entry forms, control the workflow, provide lists of possible values for each item, and check the entered values for validity. Despite this technical support, the organizational procedures around data entry have to be standardized.

27.3.1.1 Visual Editing

Since unreadable or inconsistent values impede data entry, questionnaires should be checked and edited beforehand in a step called visual editing. In case of personal interviews, the interviewer should check the data for consistency while the study participant is still present. Discrepancies, missing values, errors, or out-of-range values could be clarified on the spot. If questionnaires are collected by mail, a member of the field staff should check each questionnaire immediately after receipt. If possible, discrepancies should be clarified promptly by phone. Visual editing addresses the following aspects:

- Are entries complete and readable?
- Were more than one response option chosen although only one was allowed?
- Have skip patterns been observed correctly?

Each correction has to be documented in the questionnaire including a notice on who has done it, why, and when.

27.3.1.2 Data Entry

Guidelines describing all procedures and specifications should be laid down in a manual before data entry starts. The manual has to be handed out to the data entry staff together with hard copies of blank questionnaires and data forms. The guidelines include specifications of how to handle unreadable items and invalid values or how to use the computer system. The data entry clerk has to know whether he/she is allowed to correct the value in the questionnaire/data form or whether he/she has to enter all values just as they are written. As a true value may well fall outside predefined ranges that are based on assumptions regarding valid value ranges, the entry of implausible values should be possible. However, the entry of an implausible value should trigger an acoustic or visual warning signal.

An easy to comprehend and clearly structured manual is a prerequisite to standardize procedures between different staff members involved in data entry. Based on the guidelines, the data entry staff has to be trained on how to enter the data efficiently and accurately into the system, how to navigate through the entry system, and what should be entered (e.g., interviewer notes, comments added by the person who has done the visual editing). As part of the guideline, data entry clerks should be instructed to note his/her names (or initials) and the date of data entry on each paper questionnaire as the counterpart to the corresponding note in the database.

To motivate the data entry staff and to let them strive for maximum accuracy, they should be provided with sufficient background information about the data entry system, all screens, and the questionnaires/data forms to be entered. Additional information about the study, its goals, and context will improve the motivation, too.

The data entry system and all related procedures should be tested, including the coding and checking procedures. Such a test can best be done based on a sample of real questionnaires.

Data entry is a monotonous business carried out by a human. A data entry system can support the process, but it cannot prevent all possible errors. Errors will occur.

Table 27.3 Error proportions of data entry (first vs. second entry) for numerical variables by study center in the IDEFICS study (Taken from Ahrens et al. 2011)

	Recording sheet				
Study center	Physical examination (fasting) (%)	Physical examination (non-fasting) (%)	Parental core questionnaire (%)	Parental CEHQ[a] (%)	Physical fitness (%)
A	0.29	0.21	0.34	0.05	0.21
T	0.97	4.15	0.21	0.18	0.87
S	2.22	7.26	6.88	3.88	1.84
V	5.37	4.01	4.42	2.19	5.05
G	3.70	3.27	2.05	0.84	3.57
D	0.80	0.68	0.85	0.40	1.13
P	1.10	0.78	2.76	1.59	0.18
Z	9.42	4.36	2.58	2.96	4.85

[a] *CEHQ* Childrens's Eating Habit Questionnaire

A common approach to improve accuracy is to enter the data twice, where the second entry is preferably done by another person in a separate database. The need of (at least partial) double entry depends on reasons like the complexity of the questionnaires, the doubts concerning the correctness of data, or the required reliability and precision of data. Results from the European IDEFICS study provide a good example for the magnitude of errors occurring during data entry (Table 27.3).

Table 27.3 summarizes the comparison of first and second data entry for all numerical variables of the main questionnaires and recording sheets of the IDEFICS study. The percentage values reflect the proportion of values that differed between first and second entry averaged over all variables of the respective questionnaire/recording sheet. Further details may be found in Ahrens et al. (2011).

When the second entry is finished, the two databases can be compared, and discrepancies can be corrected. Another method is to correct the discrepancies during the second entry, where a signal indicates discrepancies. The data entry operator then instantly chooses the correct value. Reentry of only a random sample of the data may be an alternative if time and resources are limited. A tolerance limit of error frequencies should be defined prior to the comparison of both entries. If the proportion of mismatches between the first and the second entry is below the tolerance limit, a complete second entry may be considered as not necessary.

27.3.2 Data Processing

Entered data have to pass several checks and compilation procedures until they can be considered as cleaned and ready for statistical analysis (see Fig. 27.6).

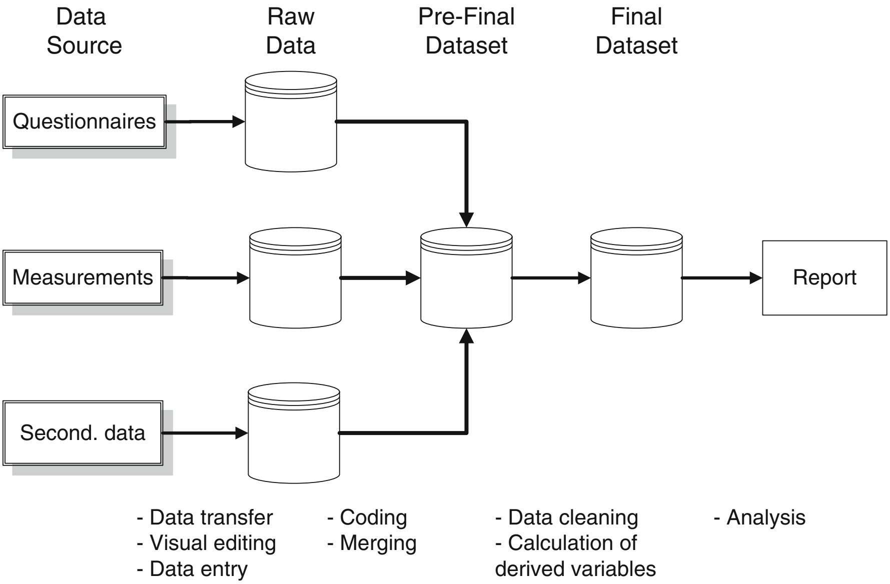


Fig. 27.6 Stages of data processing

27.3.2.1 Raw Data

The data entered from the questionnaires or collected by a measurement device are usually referred to as raw data. They reside in several data files with unequal file formats. Before the process of merging, adding, and cleaning starts, a backup copy of the raw data has to be saved. This copy is the fall-back option to restart the process in case of an error in later steps. It also serves as reference to trace back all subsequent corrections and changes.

27.3.2.2 Pre-Final Dataset

The raw data have to be linked, transformed, merged, and integrated into a uniform data structure (pre-final dataset) with uniform file formats (possibly divided into more than one data file). Codings will be harmonized and additional codes added, for example, to distinguish different sources of data.

27.3.2.3 Final Dataset

The correction of errors and inconsistencies and the exclusion of irrelevant information are part of the data-cleaning process (see Sect. 27.4), which may require continuous updates that result in a clean database for the final analyses. The enrichment of this clean database with derived variables of central importance yields the final dataset on which all statistical analyses are based. This ensures that different users work with identical data. To ease the analysis, labels and descriptions should be assigned to each variable including codes for missing values.

27.3.2.4 Processing

The necessary tasks to clean the data involve manual (by editing the data) and programmed correction routines. In contrast to manual editing, programmed routines offer the possibility to redo all changes, since these changes are documented in the program code. To give an example: In a study, social position is assigned to each participant based on an index that is composed of several variables (e.g., education, vocational training, income). In a later stage of data processing, plausibility checks show that the composed index is inaccurate. In this case the index can be recalculated with the corrected routine for all study subjects at once.

However, to benefit from programmed correction routines, each of them has to be documented, and all corrections have to be recorded in the dataset, even if an error is detected only late during data analysis. Otherwise, a correction that is only made in a specific analysis program cannot be considered in other programs. A proper documentation of correction routines and incorporation of corrections in the data ensure that all researchers always work with the latest cleaned version of the dataset.

Documentation of manual corrections should include a detailed description of which value was changed, why it was changed, and who made the change. Some database systems provide a mechanism called audit trail. If activated, all changes in the data are recorded automatically in the database. This may serve as one element to document changes made during data cleaning. A copy of the data (snapshot) should be saved after each processing step. In case of an error, the last snapshot can be used to restart the process.

If newly derived variables are generated from original variables, the original values of all source variables must be saved in addition to the new ones. Otherwise, it is impossible to recalculate the derived variables, for example, if the algorithm for their generation has to be corrected. This is particularly important if derived variables are based on outputs of measurement devices or special software products. If raw values are available, they should always be stored in addition to the resulting scores or statistics. Sometimes the manufacturer changes the algorithm of calculation in a new version of the device or software. The recalculation of the old results will be impossible without the raw data.

Generating a cleaned dataset out of the raw data is a time-consuming process, often requiring reiteration of process steps. Errors found during the data analysis can require a reimport of raw data and the repetition of various steps of the above-described process.

27.3.3 Backup and Archiving

One of the most underestimated tasks while planning and conducting a study is the development of a backup and archiving strategy for the collected data, programs, and results. Lots of efforts were undertaken to collect the data, to clean them, and to analyze them. However, everything can be lost in a few milliseconds by a hardware failure or a mistake of the user. Therefore, well-planned and carefully conducted backups are necessary. In order to set up a backup strategy, the following points have to be considered:

- Which data have to be backed up?
- Which data have to be archived?
- When should a backup be performed?

27.3.3.1 Backup of Data

It is common practice to routinely dump all files from all servers to a tape every night and keep them preserved for a certain time window. This safeguards ongoing work against hardware failures and inconsiderate maloperation. Such backup routines should be run automatically, for example, as a regular dump of all electronic data to a tape, hard disk, or DVD. Most PCs come with backup software that can also be downloaded from the internet. Such software facilitates automatic and regular backups allowing both full backups or only incremental ones without user interaction. Larger environments require a backup strategy with different levels. The choice of techniques, devices, and media (CDs, DVDs, tape archives, or hard disks) used for the backup depends on the amount of data and the requirements. In order to minimize the backup and restore time in the first level, hard disks may be used to store the backup. In the second level, the files from these disks can be written on tapes. A detailed description of possible backup scenarios can be found in De Guise (2008), Little and Chapa (2003), Nelson (2011), and Preston (2007).

Any backup mechanism has to be tested. Such a test must include a restore of some (or all) data and a check against the original data. The functionality of the backup and the restore mechanism should be tested regularly. If the data have to be restored in case of crashes, accidental deletions, or even inconsiderate disk formatting and the restore procedure does not work properly, all work is lost. The tapes or DVDs used have to be labeled with meaningful names and stored in a secure place. Old equipment and tapes should be replaced before they lose their function. It is bad practice to keep the backup on a tape or DVD drive in the computer or in the same computer rack. If the computer is destroyed, for example, by a fire, the backup will be destroyed as well. Backups must be stored in another location than the original data, preferably in a different building. Depending on the value of the data, the size of the study, and the expense needed to restore the data, additional backups may be kept in a protected place.

27.3.3.2 Archiving Data

Data have to be archived for longer time periods at the end of a project or as a snapshot of the data at the time of an interim analysis. Regular backups will be overwritten after a period of time. To archive data, another hard disk, tape, or DVD has to be used.

If the archived data have to be restored after some years, the IT environment may have changed. It cannot be excluded that archived data may not be accessible without an appropriate documentation and the programs used to store them in the first place. It is therefore good practice to archive a dataset together with its documentation (data dictionary, etc.) and any necessary files for reading and interpreting the data. If special programs (e.g., for a measurement device) use their own file formats, it must be guaranteed that these files can still be read after several years. The original program should be kept, or even better the data should be exported into a common file format (e.g., comma-separated values (csv-) file) before archiving. For example, if tables were saved as Microsoft Excel files, these files can be used for the next couple of years, also with newer versions of the program. Nevertheless, if the data were generated by a special measurement device, manufactured by a small company, the company and the device might no longer exist some years later.

The media used for archiving may be prone to ageing. Tapes, for example, could still be readable after a couple of years. DVDs should be usable too, but nobody knows exactly for how long. All the data have to be restored after a couple of years and copied to new media. Another aspect is technical innovation. Currently different types of tape drives are available, but constantly new types replace the older models. If a device is replaced by a new one, the new one must be able to read tapes used by the former one or all data have to be restored and copied to the newer models.

Many data have become inaccessible because of lost or unusable tapes or because of the unavailability of the necessary technical equipment to restore old tapes. For example, the whereabouts of data tapes of the Apollo 11 moon landing and the possibility for restoring them were unknown for a long time.[2]

27.4 Data Cleaning

The Dictionary of Epidemiology (Portas 2008) defines data cleaning "…as the process of excluding the information in incomplete, inconsistent records or irrelevant information collected in a survey or other form of epidemiological study before analysis begins. This may mean excluding information that would distort the results but it can also introduce biases. The fact that this step has been taken should be reported, along with the results of the study of analyzed data." Errors arise for a number of reasons like questionnaire faults, errors made by the interviewer, respondent misunderstanding, and (un-)intentional misreporting (response errors), laboratory and device errors, or processing errors during collection and handling of data.

Table 27.4 gives an overview of sources of erroneous (questionnaire) data in epidemiological studies. Of course, such errors should be avoided by means of well-designed, piloted questionnaires and forms, a properly designed data entry system, and well-trained and motivated staff. Anyway, errors occur in spite of such error-prevention strategies, and the process of data cleaning targets to detect, correct, or remove invalid values and records from a database. Nevertheless, data cleaning has

[2] http://en.wikipedia.org/wiki/Apollo_11_missing_tapes, last access: July 19, 2012

Table 27.4 Basic types of errors (Adapted from Van den Broeck et al. (2005))

	Type of error	
Data step	Lack or excess of data	Outliers and inconsistencies
Raw data	• Form missing/lost • Form filled in twice • Answering box left blank • Multiple answers to single choice questions	• Answer misplaced in wrong field • Illegible • Writing/typing error • Reported value outside expected range
Pre-final dataset	• Lack or excess of data carried over from questionnaire • Form or field not entered • Data erroneously entered twice • Value entered in wrong field • Inadvertent deletion or duplication of records during database handling	• Outliers and inconsistencies carried over from questionnaire • Value incorrectly entered • Data entered in wrong column of data table • Value incorrectly changed during data cleaning • Transformation (programming) error
Final dataset	• Lack or excess of data carried over from database • Data extraction or transfer error • Deletions or duplications by analyst	• Errors, outliers, or inconsistencies carried over from database • Data extraction or transfer error • Sorting errors (spreadsheets) • Data-cleaning errors

a somewhat dim existence in the context of reporting and presenting results from epidemiological studies – perhaps because it may erroneously be associated to the term data manipulation in its worst sense.

27.4.1 Steps of Data Cleaning

Usually, several people are involved in the data-cleaning process. In studies with face-to-face interviews, this process starts with the interviewers. It has early been recognized (Williams 1942) that interviewers have to follow standardized instructions exactly when using the questionnaire. They have to be neutral and to stick to the wording of questions as laid down in the questionnaire. Moreover, they should not change the order of questions in the questionnaire and always give the respondents sufficient time to make up their mind before answering. The interviewer must transfer the answers in the corresponding fields of the questionnaire with maximum care. For further details of conducting an interview, see chapter ▸Epidemiological Field Work in Population-Based Studies of this handbook. After completion of an interview, the interviewer has to cooperate with the team that is responsible for the visual inspection of questionnaires regarding missing or inconsistent information.

After questionnaires have been checked and edited, they are ready to be entered into the computer. Entry of data from written or printed documents is a potential source of errors in databases. Entering the correct value in the wrong field, transposition errors (entering a "57" instead of a "75"), and typing errors (entering a "3" instead of a "9") are the most common types of error. Fortunately, these can be detected by double data entry.

Finally, also the user of data has his/her place in the data-cleaning process by giving feedback to the data center when implausible values become apparent during data analysis. However, some basic strategies should be implemented to identify errors in a standardized way already before the analysis dataset is released.

In epidemiological studies, there is a great need of appropriate tools to clean data because manual data cleaning is laborious, time-consuming, and error prone. Therefore, it should only be applied when automatic control is impossible (e.g., correcting typos in plain texts) or uneconomical (e.g., small databases).

According to Van den Broeck et al. (2005), data cleaning can be considered as a multistage process, passing through the stages screening, diagnosing/validation, and correction, in order to identify distinctive values and to verify whether they are correct or not. One of the major problems in this process is the fact that it is not always clear whether a value is incorrect or not. Therefore, standardized algorithms on how to deal with suspect values should be developed. At this point it should be stressed that, depending on the questionnaire or the measuring instrument, there will always be some degree of deviation between measured and real values. For example, the accuracy of skinfold measurement is sensitive to the type of caliper used and also to the individual's body fat distribution. In the following we are not considering such kinds of unavoidable measurement error but only those errors that exceed small technical variations.

The following three sections highlight the three basic steps of data cleaning: data screening, diagnosing/validation, and correction of errors.

27.4.2 Data Screening

The screening of data comprises the following checks:

1. Duplicates and non-existent IDs
2. Missing or invalid values and values outside plausible value ranges
3. Inconsistencies between variables

It is advisable to proceed in this order since several inconsistencies may not occur when out-of-range checks have been made in the first place. However, detailed knowledge of the collected data and the requirements of the final analysis are needed to identify and correct wrong or inconsistent data values.

27.4.2.1 Duplicates and Non-Existent IDs

A first step in data cleaning should be to investigate whether observations occur more than once in a database. In the simplest case, a record with identical variable values occurs twice (or more). In such a case, all redundant observations have to be deleted. The situation is more complicated if a record occurs more than once and values vary between records, for example, if different data values for the same person have been entered with the correct ID several times or if the same ID has been allocated to different people. As mentioned earlier, the majority of problems with incorrect IDs, of course, may be avoided by double entry of such key variables or by use of barcodes. Anyway, for a downstream check, for example, SAS/STAT® software offers procedures like PROC SQL to search for duplicate sets of values from a predefined set of variables across two or more rows of a dataset. Cody (2008) describes such and additional very helpful procedures (PROC FREQ, PROC SORT), options (NODUPKEY, NODUPRECS), and tricks based on SAS software which automatically identify and eliminate multiple observations. These procedures are also very helpful to verify that there is a minimum number of observations per ID in a database or to check whether all IDs have exactly the same number of observations, for example, in the case of repeated measurements.

Any database should be checked for completeness, that is, whether all questionnaires and measurements from all study participants have been entered. In the ideal case, fieldwork documentation can help to check whether all paper forms, questionnaires, and electronic records of a study subject have found their way into the final analysis dataset.

27.4.2.2 Missing Values

Missing values may occur for various reasons. For example, some study participants may refuse to answer questions felt as being offensive (e.g., questions concerning income, sexual practices), or they may skip answers simply because they do not understand the question. Other respondents (or interviewers) perhaps only forget to address a question. In other situations, measurements cannot be performed for medical or ethical reasons (e.g., X-rays in pregnant women). All these situations may result in item non-response.

Missing values may also occur because a question does not apply. The answer “No” to the question “Did you ever smoke cigarettes?” should always result in a missing value or better a corresponding code for the subsequent question that is legitimately skipped (see below): “How many cigarettes did you smoke per day?” Conversely, if the number of cigarettes smoked per day has been reported but the answer to the previous question whether the respondent ever smoked cigarettes is missing, then the answer can be imputed as “Yes” ex post during data cleaning.

Different codes should be used to distinguish between item non-response and missing values for other reasons to allow quality checks of their respective frequency. As mentioned above, it is recommended to use values that are out of range of the valid values of a given variable to represent different types of missing values (see Table 27.1). Nevertheless, whenever possible, missing information may be replaced by valid values from available sources. For example, a missing value for the variables sex or date of birth in a longitudinal study may be substituted by the corresponding information obtained during follow-up.

If there is no relationship between item non-response and the values of other variables, these missing values are called missing completely at random (MCAR). If the probability of missing data depends on other observed variables, this is a systematic error. However, when the missing values are at random after controlling for these other observed variables, that is, when they are conditionally random, such missing values are called missing at random (MAR). Finally, if the missing values are not completely random and may not be completely explained by other variables in the dataset, then such missings are considered as not missing at random (NMAR), which means that the probability that values are missing depends on other variables of interest or on the variable itself (e.g., the probability that self-reported survey data on weight is missing is higher for overweight participants of a study). Whenever it is likely that data are not missing at random, it is important "to carefully assess the sensitivity of results to a variety of plausible assumptions concerning the missingness process" (Fitzmaurice 2008). Many different methods have been proposed to handle missing data during statistical analysis. A comprehensive overview may be found in Sterne et al. (2009) and in chapter ▸Missing Data of this handbook.

27.4.2.3 Implausible Values

The search for incorrect or implausible data values starts with a univariate, descriptive analysis of single variables. For identifying data errors, one examines the distribution of a variable, possibly using a graphical approach, but in any case calculating measures of location (quartiles, mean, median, minimum, and maximum) and variation (standard deviation or variance). In this context, descriptive statistics are useful tools for explorative data analysis and especially for the detection of suspected values.

Often it is difficult to identify implausible values only by means of univariate methods, that is, without considering a broader context. Some data values may not catch one's eye before they are combined with other variables, that is, certain values only fall outside an expected range when considered in combination with another variable. Such a bivariate inspection of data may also be considered as a consistency check (see below). Let us assume that a child's true body weight of 18 kg was erroneously entered as 10 kg. Data cleaning should find ways to identify such errors. In this example, a useful plausibility check would combine the variable body weight with height, for example, by means of a scatterplot as a simple tool for spotting outliers. In a scatterplot, each subject is presented as a point on the x- and y-axis. Most subjects will fall within the cloud of points, while outliers tend to be located outside the cloud. Figure 27.7 shows such a plot of height and weight of 2- to 9-year-old children. The weight value of a child in the above example becomes suspect with the unusual combination of 10 kg and a height of 118 cm.

Nevertheless, it can be easily conceived that such an approach is not always successful. If, in the above-mentioned example, the body weight has been erroneously entered as 28 kg, the mistake remains undetected.

Boxplots are another approach to detect and visualize implausible data values (Tukey 1977). For a continuous variable a boxplot summarizes several statistical measures like mean, median, upper and lower quartiles, as well as minimum and maximum data values in a graph which displays a variable's location, spread, and

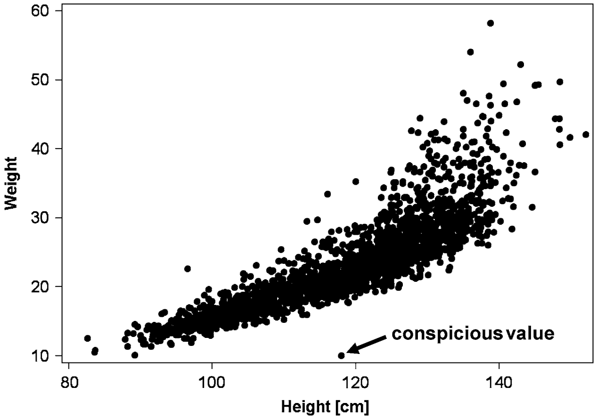


Fig. 27.7 Scatterplot of height and weight of 2- to 9-year-old children

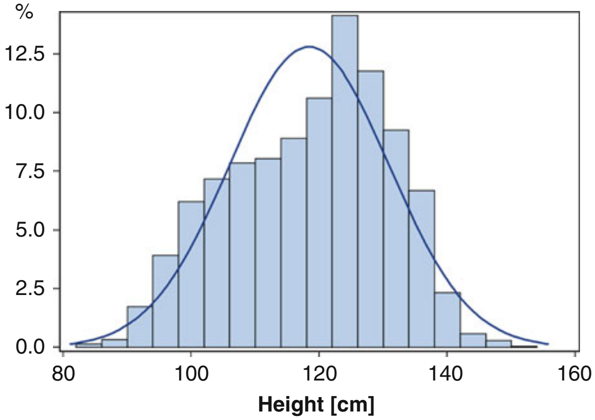


Fig. 27.8 Histogram and density (height)

outliers at a glance. Boxplots may be stratified by a second correlated categorical variable to detect outliers. The univariate distribution of the variable "height" of 2- to 9-year-old children is shown in Fig. 27.8. This graphical presentation shows no abnormalities in terms of implausible values. Only when stratified by age the side-by-side boxplots reveal several values of body height as suspect (Fig. 27.9).

Table 27.5 gives an overview for the variable height regarding the most important measures of location and dispersion as well as reasonable age-specific ranges. From this table it may be seen that the observed range for 2- to 3-year-old children is completely covered by the corresponding expected ranges, which contrasts to the observed range in the group of 5- to 6-year-old children.

27.4.2.4 Consistency Checks

Data combinations that are logically impossible may occur when data have been entered correctly but when the interviewer or the study participant made a mistake when filling in the questionnaire. Such errors can be identified by finding impossible

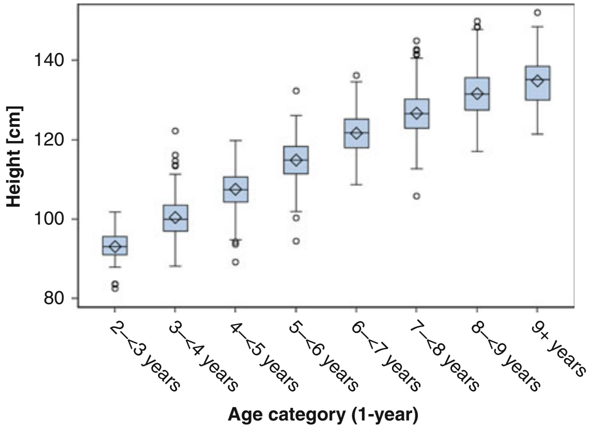


Fig. 27.9 Age-specific boxplots for height

Table 27.5 Descriptive statistics of height by age of children (Hebestreit and Ahrens 2012)

Summary descriptive table	Expected (“reasonable”) values of height[a]	Number of children	Height (cm)						
			Mean	SD	Min	Q1	Median	Q3	Max
Age group									
2–<3 years	79–103 cm	46	93.2	4.2	82.6	91.1	93.2	95.6	101.8
3–<4 years	87–111 cm	255	100.5	4.9	88.2	97.0	100.0	103.5	122.3
4–<5 years	93–118 cm	337	107.5	4.9	89.2	104.3	107.4	110.7	119.8
5–<6 years	99–125 cm	238	114.9	5.2	94.5	111.5	115.0	118.4	132.4
6–<7 years	106–133 cm	333	121.7	5.2	108.7	118.0	121.8	125.2	136.2
7–<8 years	112–140 cm	476	126.7	5.8	105.9	123.0	126.6	130.3	145.0
8–<9 years	117–146 cm	334	131.7	5.8	117.1	127.5	131.5	135.7	149.9
9–<10 years	121–152 cm	46	134.9	6.4	121.5	130.1	135.3	138.6	152.1
All		2,065	118.4	12.5	82.6	108.5	120.0	128.0	152.1

[a] Expected values of height are based on the 3th and 97th percentile published by Kuczmarski et al. (2000)

or extremely unusual combinations of variable values such as male patients who reported a Caesarean section, female patients with prostate cancer, or if the date of admission to a hospital precedes the subject’s date of birth. Unfortunately, it is nearly impossible to create consistency checks for all conceivable combinations of variable values in a dataset. Many inconsistencies are only revealed during data analysis. In any case, it has to be documented how observed inconsistencies are treated.

If several questionnaires are used within a study or variables are recorded repeatedly (e.g., in follow-ups), one should check whether specific characteristics such as date of birth and sex are consistent. Inconsistencies between such repeatedly collected variables might suggest that the forms belong to different respondents.

27.4.3 Data Validation and Correction of Errors

Data validation and correction of errors means the process of checking information in the database for consistency and plausibility in order to identify errors and correct them. At this point we are not referring to the validation and correction of first and second data entry or first and second codings (e.g., for job tasks, branches of industries, or diseases) – such discrepancies should be corrected considering the original questionnaire data or discussed and solved by the coding experts at regular team meetings through consensus. Here we are thinking of reviewing the data against established criteria in order to accept or reject suspect values. One opportunity for validation is the revision or confirmation of data values using external sources, when, for example, researchers review medical records to confirm the occurrence of events reported by patients.

Checks of value ranges serve to identify outliers, that is, values of a variable that fall outside the range of reasonable values for that variable. For example, a percentage measurement must be in the range of 0 and 100, or the height of an adult person must be in the range of 50–250 cm. Therefore, if you detect values above 250 cm, it is obvious that an error was made, either by filling the questionnaire or during the data entry process. Of course, range checks should be first of all part of data entry to allow immediate cross-checks with the data source (cf. Sect. 27.3.1) or to accept only responses within a specified range. In the absence of reference data or obvious normal ranges, it becomes difficult to define the upper and lower limits of observed values. In these cases, objective and systematic rules for the detection of outliers, that is, observations that deviate "too much" from other observations, are needed. As a simple rule of thumb, (Osborne 2010) suggested to consider observations that deviate by more than three standard deviations (SDs) from the mean of a normally distributed variable as "not generated by the population of interest" (Osborne 2010). In any case, one should be very cautious before correcting or removing an outlier. The deletion of an outlier leads to a smaller "new" standard deviation. As a consequence now, other previously unsuspected values may appear as outliers. Figure 27.10 shows the distribution of data and possible cut-offs to distinguish between questionable and impossible values. However, in the case of non-normal distributions, data have to be normalized by appropriate transformations before looking for outliers.

According to Van den Broeck et al. (2005), there are three options to treat a suspect value: correcting, deleting, or leaving unchanged. Close attention should be paid to the question at which stage of data processing one carries out these changes. In general, the raw dataset should never be changed, perhaps with one exception: a mismatch between the information on the questionnaire and the corresponding value in the data file; such discrepancies should be corrected immediately in the raw dataset. All other types of errors and inconsistencies should be identified and removed during the data-cleaning process. Impossible values should be set to missing or changed to a value (e.g., −9) that indicates that the original value is incorrect. Unchanged suspect values may be flagged. Flagged suspect values can be easily excluded from subsequent sensitivity analyses. If the reason for a false

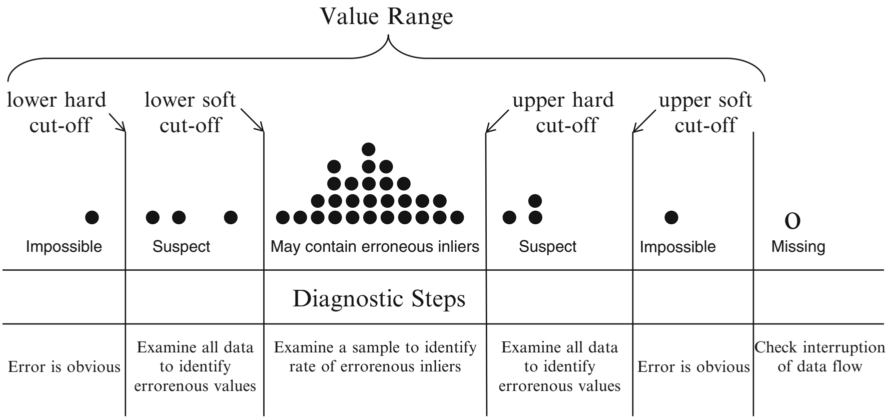

Value Range
lower hard cut-off
lower soft cut-off
upper hard cut-off
upper soft cut-off
Impossible
Suspect
May contain erroneous inliers
Suspect
Impossible
O
Missing
Diagnostic Steps
Error is obvious
Examine all data to identify errorenous values
Examine a sample to identify rate of errorenous inliers
Examine all data to identify errorenous values
Error is obvious
Check interruption of data flow

Fig. 27.10 Hard and soft cut-offs in data cleaning (Adapted from Van den Broeck et al. (2005))

value can be traced back, it may be justified to use other variables to reconstruct the correct value and to replace the suspect value by a plausible value (cf. Sect. 27.4.3). However, often information is insufficient to clarify suspect values. Then removal of the questionable values may be the only choice.

In general, automatic corrections should only be made in unambiguous situations (see the example in Table 27.6 where the respondents mixed up the week of delivery with the weeks before the calculated delivery date). Suspect values that cannot be corrected by a standardized routine should be clarified case by case by experienced researcher staff that is able to define hard cut-offs or to decide whether an extreme value is still (biologically) possible. Hard cut-offs should always be set in a conservative way in order to minimize the chance that a suspect value is deleted although it is true.

By means of simple analysis procedures in statistical software packages, it is possible to produce the so-called exception reports, sorted by ID number. An exception report is a listing of suspect values that have become conspicuous in the course of data screening. Since such values in general fall outside a predefined range, a listing of additional information may be helpful for assessing the implausibility (triangulation). For example, if the weight of a child seems unusually high, it may be helpful to list further additional information such as height, waist circumference, or skinfold thickness. Different approaches for creating such lists by means of the SAS program package can be found in Cody (2008).

Table 27.6 shows an excerpt from such an exception report. As a first check of suspect values, data entry errors should be ruled out by comparison with the original data source. If a suspect value roots in the original documents or the original data files, it may be useful to clear it with the staff involved in data collection or data entry. In general, such mutual consultations between interviewers, study nurses, data typists, physicians, statisticians, and epidemiologists involved in the study may help to decide difficult cases and to settle decision rules. All principal decisions regarding

Table 27.6 Example for an exception report taken from the IDEFICS study (Ahrens et al. 2011)

ID	Question: variable	Description	Value	Confirm or correct
1122	Q1: Child's date of birth	Out of range (01.01.1999; 31.12.2005)	25.04.2007	25.04.2000
3243	Q3: Weeks before calculated date (weeks)	Weeks before calculated date >> upper limit (6 weeks)	34	6
5243	Q4: Birth – child's weight (g)	Child's birth weight << lower limit (1,400 g)	1,380	✓
1467	Q5: Birth – child's height (cm)	Child's birth height << lower limit (45 cm)	44	✓
1499	Q2: What sex is the child?	Different information from T0 and T1	Male	Male ✓

the data-cleaning process should be agreed with the staff that is responsible for data management. Corrections should be made on the listing first (last column of Table 27.6) and in the data file subsequently.

In this example, the child's date of birth for ID 1122 was mistyped, and the correct date as given in the questionnaire was entered manually in the last column of the list. The birth weight of ID 5243 as well as the birth height of ID 1467 was confirmed by inspection of the original questionnaire. Often it is useful to complete such a list with additional information that helps to verify questionable values. For instance, it can be useful to list the complete questionnaire information on preterm delivery, birth weight, and height simultaneously to facilitate the validation of suspect values at a glance. The value of 34 weeks for preterm birth of ID 3243 was correctly transferred from the questionnaire. Obviously parents completing the questionnaire misunderstood the question in this case and reported the week of delivery instead of the weeks before the calculated date of delivery. In such obvious situations it may be justified to recalculate the value in question. Here, the difference between the full term and the reported week of delivery has to be calculated as $40 - 34 = 6$ weeks before calculated date. This example shows that it may be justified to perform corrections automatically. Here the following correction algorithm was implemented: if reported value $x > 30$, then calculate the corrected value as $x' = 40 - x$. For ID 1499, contradicting information was noted for sex of the child reported at baseline (T0) and at follow-up (T1). Depending on the study protocol, such disagreements may be set to missing or corrected (in case there is an additional trustworthy data source), or the original respondent may be contacted for verification.

It is advisable to carry out the whole process of data cleaning within a command file. This ensures a complete documentation of changes made to a dataset (see Table 27.7).

Table 27.7 Example for an SAS-based command file for the correction of errors

```
Data libname.final_data;
    set libname.pre_final_data;
              ...
                If id_no=1122 then year_of_birth=2000;
                If id_no=1499 then sex_child=1;
                If weeks_before>30 then weeks_before=40-weeks_before;
            ...
            ...
```

Again, it is strongly recommended to never change the raw data and to store all programs that have been used at different time points to create the final dataset. This ensures that also intermediate versions of the database can be reconstructed. After the correction of errors, the above-described methods of data screening should be repeated to make sure that data that are out of range of reasonable limits have been corrected.

27.5 Preparing the Final Dataset

A study protocol of an epidemiological study describes the hypotheses and the means by which these are investigated and tested. On an operational level, precise definitions of the study variables and a description of the planned statistical analyses are necessary. This may include a description of the statistical models to be used, a list of variables to be included in the models, the specification of a variable selection procedure, and the planned subgroup and sensitivity analyses. In preparing these proceedings, one can learn from the conduct of clinical trials, where the consent of regulatory authorities is obtained on the basis of definitions and operational procedures laid down in a so-called Statistical Analysis Plan (SAP). An outline of such plan is given in Box 27.1.

To allow for new questions and hypotheses that emerge in course of a study, an SAP may be updated where revisions are numbered and each version is documented.

27.5.1 Final Dataset

Statistical analyses should be performed on a prepared analysis dataset (Chang and Wong 2005), which we call the final dataset. It contains all variables, original and derived, on which the planned final calculations are performed. In this way, the final dataset serves as a repository of all variables on which the final statistical outputs are based. Separating the final dataset from data management processes like merging, cleaning, or creating derived variables has several advantages: (i) reproducibility – all steps leading from raw data to statistical outputs can easily be reproduced;

Box 27.1. Outline of a Statistical Analysis Plan (SAP)

Suggested outline of an SAP:

(i) *Introduction*: background and aim(s) of the study

(ii) *Data management*: methods for the collection and use of data
   (a) Data flow
   (b) Analysis datasets
      – Inclusion/exclusion criteria for participants
      – Variables (original and derived) to be included for the analyses
      – Definition of derived variables (see below)
      – Handling of missing values and outliers
      – Matching strategies (if applicable)

(iii) *Statistical analyses*: planned statistical evaluations
   (a) Descriptive analyses
   (b) Primary analyses, with full model specification (covariates, reference categories, etc.)
   (c) Statistical methods, if these are not standard
   (d) Subgroup and sensitivity analyses and power calculations

(iv) Measures to assure quality
   (a) Program documentation and validation

(ii) comprehensibility – distinct steps of data processing are clearly separated; and (iii) simplicity – the analysis programs become more simple. For studies incorporating a wide variety of potential risk factors, the final dataset may be structured according to different areas, for example, demography, lifestyle, and occupational exposure.

27.5.2 Quantification of Exposure

Operationalizing "exposure," be it chemical agents in a working environment or personal habits like physical exercise or smoking, is almost always a substantive task requiring interdisciplinary collaboration, as it is extensively documented in this handbook. This task requires considerably more person-time than the final statistical analysis. The process of transforming raw data, comprising the information collected for each participant, into numerical variables that can be used in statistical procedures is called quantification.

Definitions and computational details for derived variables should be laid down in the SAP or/and a similar document (data dictionary, see Table 27.2 for present and later reference). When quantifying the exposure, the first step is to choose an appropriate metric; often, more than one metric will be derived (see chapter ►Exposure Assessment of this handbook). In Table 27.8 four examples are given, starting with quantifying demographic variables.

Table 27.8 Example of a calculation table describing study variables of an analysis dataset

Variable name	Description	Details
	Physical activity	
PHYS_ACTIV	Hours of physical activity per week	*PHYS_ACTIV = 5*(PLWEEKD_H + PLWEEKD _M/60) + 2*(PLWEEKEN_H + PLWEEKEN_M/60) + (CLUB_H + CLUB_M/60)*
	Demographic variables	
YEAR	Year of interview	e.g., 2009
AGE	Age at cohort entry in years calculated by "entry year – birth year"	Positive integer, e.g., 56
AGE_CAT	Categorization of *AGE*	1: $AGE \leq 40$ 2: $40 < AGE \leq 50$ 3: . . . . .
AGE_C	*AGE* centered at 55 years	*AGE_C = AGE – 55*
	Smoking	
PACKY	Pack-years (sum over smoking periods (in years) times number of smoked packages per day)	Data from smoking biography, assuming *J* smoking periods, each starting with *YBEG*, ending with *YEND*, with *PACK* (per day) $PACKY = \sum_j (YEND_j - YBEG_j + 1)*PACK_j$
PACK_CAT	Categorization of *PACKY*	0: Never smoked cigarettes $1: 0 < PACKY \leq 10$ $2: 10 < PACKY \leq 20$ $3: 20 < PACKY \leq 40$ $4: PACKY > 40$
LOG_PY	Log(pack-years + 1)	*LOG_PY*=log(*PACKY*+*1*)
SMOKE_DUR	Duration of smoking	$\sum_j (YEND_j - YBEG_j + 1)$
SMOKER	Person has smoked cigarettes, cigars, or pipes	=1, if (*PACKY*>½ or person has smoked cigar or pipe for more than 1 year) =0 otherwise
	Occupation	
ASBESTOS	Self-reported occupational asbestos exposure	0: No 1: Yes
List_A_Dur	Duration in industries/occupations known to be associated with lung cancer (for the list of corresponding codes, see Ahrens and Merletti 1998)	Data from job periods, assuming *J* periods of occupations that belong to List A, each starting with JOB_*BEG*, ending with JOB_*END* $List\_A\_Dur = \sum_j (JOB\_END_j - JOB\_BEG_j + 1)$

The second example comes from an ongoing cohort study (Ahrens et al. 2011) and describes measurement of a child's weekly physical activity. Total "hours of physical activity per week" may be expressed as the total time per week performing physical activity including outdoor activities and the time spent in a sports club.

The derived variable for this summary measure, *PHYS_ACTIV*, is calculated from the (questionnaire) variables *PLWEEKD_H* and *PLWEEKD_M* (hours and minutes the child was playing outdoors on weekdays), the variables *PLWEEKEN_H* and *PLWEEKEN_M* (hours and minutes the child was playing outdoors on weekend days), and the variables *CLUB_H* and *CLUB_M* (hours and minutes a child spent in a sports club per week). Single missing values are set to zero. The exact quantification of this variable can be seen in Table 27.8.

The third example describes quantification of smoking data from a case-control study where participants reported a detailed smoking history. The interview assessed changes of smoking behavior over time and recorded type and amount smoked for each period over which smoking behavior was stable. The following variables are available for each smoking period: *YBEG* (first year of smoking period), *YEND* (last year of smoking period), and *PACK* (packages (20 cigarettes) per day). Cigarette smoking is often quantified in terms of a cumulative measure called pack-years (see chapter ▸Environmental Epidemiology of this handbook). This metric combines duration and intensity of cigarette smoking by multiplying the number of smoked packages per day with duration in years. Further variables like smoking categories or a log-transformed variable log(pack-years + 1) can be derived from this metric. Other frequently used exposure variables are ever/never exposed (*SMOKER*) and duration of exposure in years (*SMOKE_DUR*).

Also based on the data from a case-control study, the fourth example describes the quantification of occupational hazards. To evaluate the total impact of occupational exposure on lung cancer, Pohlabeln et al. (2000) used the work of Ahrens and Merletti (1998) which translated occupations and industries that are known (List A) or suspected (List B) to be associated with lung cancer into the corresponding ISCO (International Labour Office 1968) and/or ISIC codes (United Nations Publications 1971). Based on these codes it was possible to calculate, for example, the duration of employment in List A occupations.

Sometimes a transformation of variable values (e.g., logarithm, square root) is necessary to accommodate the assumptions underlying a statistical model (e.g., normality, constant variance) or the assumed biological dose-effect relationship. The three main reasons for transforming data are (i) to change their distributional form (usually to make it closer to a normal distribution), (ii) to stabilize variance, and (iii) to linearize relationships (Armitage et al. 2002). For some well-investigated factors like smoking, certain transformations of the corresponding metric were shown to be useful. For example, the previously mentioned log-transformed variable log(pack-years + 1) often produces a better fit in regression models compared to the untransformed variable (Breslow and Day 1980). For further methods of fitting continuous variables for statistical modeling, for instance, building polynomial expressions, see chapter ▸Analysis of Continuous Covariates and Dose-Effect Analysis of this handbook).

Three aspects should be considered when variables are transformed or otherwise fitted for statistical modeling: (i) quality assurance, that is, documentation and validation (e.g., in the form of program code review or even by double programming) of the program code, is of utmost importance; (ii) transparency and clarity have to be enhanced by labeling of variables and by using formats for categorical variables; and (iii) full precision of the recorded data (no "rounding" of intermediate results) has to be used for calculations.

Quantification may also involve the application of complex statistical concepts and procedures like, for instance, a principal component analysis to condense the number of variables of a food frequency questionnaire or propensity score methodology in a pharmacoepidemiological study. For example, in an investigation on adverse events related to the use of a drug, a propensity score adjusted analysis is foreseen. Calculating the propensity score itself relies on prior quantification of co-medication and comorbidities thought to be related to drug prescription. Prior quantification is also necessary for simpler measures like the Charlson comorbidity score (Sax and Charlson 1987) that is also employed in the field of pharmacoepidemiology.

A different type of quantification concerns the integration of extraneous information to enhance exposure assessment, as, for example, with job exposure matrices (JEMs) in occupational epidemiology (see chapter ▸Occupational Epidemiology of this handbook). Here the coded occupations experienced during a subject's job history are linked to a JEM to infer agent exposures which are inferred from job titles. An even more thorough (and more expensive) assessment would make the complete working biography accessible to industrial hygienists to perform an individual exposure rating (Pohlabeln et al. 2002, see also chapter ▸Exposure Assessment of this handbook).

Finally, two major considerations should be kept in mind when deriving study variables:

(i) Time-related exposure variables have to be truncated prior to disease onset of cases and a corresponding point in time of referents to avoid misinterpretation (e.g., reverse causation). In chronic disease epidemiology the time of disease onset usually is unknown as is the induction period of an exposure to a suspected agent, that is, the time between exposure and its effect in form of clinical disease manifestation. It is therefore prudent to restrict exposure duration to a period that ends a considerable time prior to interview or diagnosis. The approach may be formalized by computing time-weighted cumulative exposure measures (see chapter ▸Environmental Epidemiology of this handbook or Breslow and Day (1987)). As an example of what might go wrong, consider a case-control study on risk factors of leukemia. Diagnostic X-rays are discussed as a potential risk factor of leukemia and lymphoma (Boice Jr. et al. 1991). If – in a case-control study – assessment of X-ray exposure would include the time period of disease manifestation, the history of cases would include all the diagnostic X-ray exposures caused by diagnostics to localize the disease. As non-diseased persons (controls) would usually not experience this exposure, reverse causation might occur, that is, X-ray exposures are the result of the diagnosis (not the other way round).

(ii) With secondary data, often outcome variables have to be constructed. In epidemiological field studies, disease status is usually observed directly, and great scrutiny is exercised in ascertaining the case status of a patient. For example, in studies of lung cancer, a diagnosis has to be histologically confirmed by tissue material obtained via bronchoscopy or surgery. When working with secondary data like claims data of statutory health insurances, outcome variables like disease status are not observed directly. Therefore, proxy indicators of the disease status of a subject have to be derived out of several choices. Here, high specificity deserves priority; otherwise, the case series would include a large number of false positives, leading to biased relative risk estimates. The following may serve as an example: Based on claims data, an investigator wants to assess the impact of certain medical risk factors on the onset of gastrointestinal bleeding (GIB). The diagnosis of a GIB event may be recorded, both, in ambulatory or hospital data. In this case it is recommended to use hospital records only, specifically the discharge diagnosis, being the most reliable one.

27.6 Conclusions

Data management in epidemiological studies includes the collection, entry, and quality control of data as well as the preparation of a cleaned database. One of the basic prerequisites for the provision of a high-quality database is, of course, that it has to be ensured that field staff adheres to standardized instructions in using questionnaires and measurement devices. Any automatical transfer of data from the devices to the database should ensure that the actually measured values are transferred. Dealing with missing values, the criteria for the definition of implausible values and the reasons which lead to the exclusion of entire records, as well as the quantification and operationalization of derived variables, must be described comprehensibly in the study manual.

According to Chin and Lee (2008), an error detection arsenal consists of five major tools: visual inspection, data entry programs, descriptive statistics, graphing, and logical checks. Data collected by questionnaires should always be visually inspected and entered twice (double data entry). In addition, data entry should be technically restricted to plausible ranges and valid codes (for alphanumerical variables). Immediately after entering the first few records, one should start the process of checking the incoming data for completeness, plausibility, and consistency. This can be done by simple descriptive statistics (minimum, maximum, frequency of missing values), possibly supplemented by graphical methods (e.g., boxplots). Stratification by selected variables (e.g., age or country) as well as multidimensional graphical methods (e.g., scatterplots) may help to identify implausible combinations of values. Conspicuous values then should be identified by a program and printed out as lists (exception report) which may be supplemented by further variables that can be helpful in assessing the plausibility (triangulation).

Corrections in the original questionnaire and the raw dataset should be avoided, perhaps with one exception: If it can be proved beyond doubt that the collected information is really incorrect, then the error should be corrected in the original questionnaire as well as in the raw dataset – accompanied, of course, by a corresponding documentation in the questionnaire. All other corrections should be implemented by a computer program which leaves the raw data files unchanged but generates a so-called pre-final dataset. This pre-final dataset should ideally only comprise variables with plausible and valid values, whereas, of course, conspicuous values that were validated and found to be correct (e.g., if a child indeed weighs 120 kg) remain unchanged.

Data entry and verification of the data should be done as soon as possible to identify and resolve any emerging systematic errors, caused, for example, by faulty devices. Based on the verified and corrected pre-final dataset, the final dataset (analysis dataset) will be generated by means of a well-documented computer program. This database contains all variables, original and derived, on which the final calculations are performed. In this way, the final dataset serves as a repository of all variables on which the final statistical outputs are based.

Chin and Lee (2008) made a nice comparison between the process of data management and cooking: "Before cooking, a good chef checks the quality of the ingredients (e.g., how fresh are the vegetables and fish, what is the grade of the meat) and then carefully converts the ingredients into usable forms (e.g., removes the bones from the fish, dices the vegetables, and tenderizes the meat). Without good quality ingredients, no chef [. . .] can prepare good quality dishes. Without properly preparing the ingredients, the dishes also will suffer." Comparable to that, ". . . poor quality data can only yield poor quality results that can only generate very limited conclusions."

References

Ahrens W, Merletti F (1998) A standard tool for the analysis of occupational lung cancer in epidemiologic studies. Int J Occup Environ Health 4:236–240

Ahrens W, Bammann K, Siani A, Buchecker K, De Henauw S, Iacoviello L, Hebestreit A, Krogh V, Lissner L, Mårild S, Molnár D, Moreno LA, Pitsiladis YP, Reisch L, Tornaritis M, Veidebaum T, Pigeot I; IDEFICS Consortium (2011) The IDEFICS cohort: design, characteristics and participation in the baseline survey. Int J Obes 35(Suppl 1):S3–15

Armitage P, Berry G, Matthews JNS (2002) Statistical methods in medical research, 4th edn. Blackwell, Oxford

Boice JD Jr, Morin MM, Glass AG, Friedman GD, Stovall M, Hoover RN, Fraumeni JF Jr (1991) Diagnostic x-ray procedures and risk of leukemia, lymphoma, and multiple myeloma. JAMA 265:1290–1294

Breslow NE, Day NE (1980) Statistical methods in cancer research. Volume I – the analysis of case-control studies. IARC Science Publication, Lyon

Breslow NE, Day NE (1987) Statistical methods in cancer research. Volume II – the design and analysis of cohort studies. IARC Science Publication, Lyon

CDC Centers for Disease Control and Prevention (2011) Epi Info™ 7. http://www.cdc.gov/epiinfo/. Accessed 9 Aug 2012

Chang S, Wong S (2005) The role of analysis datasets in successful FDA advisory meetings. http://www.lexjansen.com/pharmasug/2005/fdacompliance/fc06.pdf. Accessed 9 Aug 2012

Chin R, Lee B (2008) Principles and practice of clinical trial medicine. Academic, St. Louis

CIOMS (2008) International ethical guidelines for epidemiological studies. Council for International Organizations of Medical Sciences (CIOMS) in collaboration with the World Health Organization (WHO), Geneva

Cody R (2008) Cody's data cleaning techniques, 2nd edn. SAS Institute Inc. Cary, NC

Dean AG, Sullivan KM, Soe MM (2011) OpenEpi: open source epidemiologic statistics for public health, Version 2.3.1. http://www.openepi.com/OE2.3/Menu/OpenEpiMenu.htm. Accessed 12 July 2012

De Guise P (2008) Enterprise systems backup and recovery: a corporate insurance policy. Auerbach Publications, Boston

EpiData Association (2010) EpiData Software. http://www.epidata.dk/. Accessed 9 Aug 2012

Fitzmaurice G (2008) Missing data: implications for analysis. Nutrition 24:200–202

Gassman JJ, Owen WW, Kuntz TE, Martin JP, Amoroso WP (1995) Data quality assurance, monitoring, and reporting. Control Clin Trials 16:104S–136S

Gumm HP (1986) Encoding of numbers to detect typing errors. Int J Appl Eng Educ 2:61–65

Hartge P (2006) Participation in population studies. Epidemiology 17:252–254

Hebestreit A, Ahrens W (2012) Dietary and lifestyle-induced diseases in children: design, examination modules and study population of the baseline survey of the German IDEFICS cohort (in German). Bundesgesundheitsblatt 55:892–899

International Labour Office (1968) International standard classification of occupations. International Labour Office Publications, Geneva

IEA International Epidemiological Association (2007) Good Epidemiological Practice (GEP) IEA Guidelines for proper conduct in epidemiological research. http://www.iaeweb.org/. Accessed 11 May 2012

Kuczmarski RJ, Ogden CL, Grummer-Strawn LM, Flegal KM, Guo SS, Wei R, Mei Z, Curtin LR, Roche AF, Johnson CL (2000) CDC growth charts: United States. Advance data from vital and health statistics. National Center for Health Statistics, Hyattsville

Likert R (1932) A technique for the measurement of attitudes. Arch Psychol 140:1–55

Little DB, Chapa DA (2003) Implementing backup and recovery: the readiness guide for the enterprise. Wiley, Indianapolis

Morton LM, Cahill J, Hartge P (2006) Reporting participation in epidemiologic studies: a survey of practice. Am J Epidemiol 163:197–203

Nelson S (2011) Pro data backup and recovery. Apress, New York

Olson SH, Voigt LF, Begg CB, Weiss NS (2002) Reporting participation in case-control studies. Epidemiology 13:123–126

Osborne JW (2010) Data cleaning basics: best practices in dealing with extreme scores. Newborn Infant Nurs Rev 10:37–43

Pohlabeln H, Boffetta P, Ahrens W, Merletti F, Agudo A, Benhamou E, Benhamou S, Brüske-Hohlfeld I, Ferro G, Fortes C, Kreuzer M, Mendes A, Nyberg F, Pershagen G, Saracci R, Schmid G, Siemiatycki J, Simonato L, Whitley E, Wichmann HE, Winck C, Zambon P, Jöckel KH (2000) Occupational risks for lung cancer among nonsmokers. Epidemiology 11:532–538

Pohlabeln H, Wild P, Schill W, Ahrens W, Jahn I, Bolm-Audorff U, Jöckel KH (2002) Asbestos fibreyears and lung cancer: a two phase case-control study with expert exposure assessment. Occup Environ Med 59:410–414

Portas M (2008) A dictionary of epidemiology. Oxford University Press, New York

Preston CW (2007) Backup & recovery: inexpensive backup solutions for open systems. O'Reilly Media, Sebastopol

Prud'homme GJ, Canner PL, Cutler JA (1989) Quality assurance and monitoring in the Hypertension Prevention Trial. Control Clin Trials 10:84S–94S

Reineke A, Pigeot I, Ahrens W (2014) MODYS – a modular control and documentation system for epidemiological studies. In: Bammann K, Ahrens W (eds) Instruments for a large sacle survey in children – the European IDEFICS study: development, scientific rationale, application and practical recommendations. Springer, Heidelberg

Sax FL, Charlson ME (1987) Medical patients at high risk for catastrophic deterioration. Crit Care Med 15:510–515

Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 338:b2393

Theobald K, Capan M, Herbold M, Schinzel S, Hundt F (2009) Quality assurance in non-interventional studies. Ger Med Sci (GMS) 7:Doc29

Tooth L, Ware R, Bain C, Purdie DM, Dobson A (2005) Quality of reporting of observational longitudinal research. Am J Epidemiol 161:280–288

TrueCrypt (2012) Free open-source on-the-fly encryption. http://www.truecrypt.org/. Accessed 12 July 2012

Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading

United Nations Publications (1971) International standard industrial classification of all economic activities (ISIC). Publishing Service United Nations, New York

Van den Broeck J, Cunningham SA, Eeckels R, Herbst K (2005) Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med 2:e267

van Es GA (1996) Research practice and data management. Neth J Med 48:38–44

Vrijheid M, Richardson L, Armstrong BK, Auvinen A, Berg G, Carroll M, Chetrit A, Deltour I, Feychting M, Giles GG, Hours M, Iavarone I, Lagorio S, Lonn S, McBride M, Parent ME, Sadetzki S, Salminen T, Sanchez M, Schlehofer B, Schuz J, Siemiatycki J, Tynes T, Woodward A, Yamaguchi N, Cardis E (2009) Quantifying the impact of selection bias caused by nonparticipation in a case-control study of mobile phone use. Ann Epidemiol 19:33–41

Whitney CW, Lind BK, Wahl PW (1998) Quality assurance and quality control in longitudinal studies. Epidemiol Rev 20:71–80

Wichmann H-E, Kaaks R, Hoffmann W, Jöckel K-H, Greiser KH, Linseisen J (2012) The national cohort (in German). Bundesgesundheitsblatt 55:781–789. see also: http://www.nationale-kohorte.de. Accessed 9 Aug 2012

Williams D (1942) Basic instructions for interviewers. Public Opin Q 6:634–641

World Health Organization (2009) International statistical classification of diseases and health related problems. The ICD-10, 2nd edn. World Health Organization, Geneva

Sample Size Determination in Epidemiological Studies

28

Janet D. Elashoff and Stanley Lemeshow

Contents

J.D. Elashoff (✉)
Statistical Solutions, Saugus, MA, USA

S. Lemeshow
Dean, College of Public Health, The Ohio State University, Columbus, OH, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_15,

28.1 Introduction

When planning a research project, an epidemiologist must consider how many subjects should be studied. While factors such as available budget certainly present constraints on the maximum number of subjects that might actually be included in a study, statistical considerations are extremely important. To address the statistical questions about appropriate sample size, the researcher must first specify the study design, the nature of the outcome variable, the aims of the study, the planned analysis method, and the expected results of the study. Is the goal of the study to distinguish between hypotheses about the value of a parameter or function of parameters, or is the goal to provide a confidence interval estimate of a parameter such as the odds ratio or relative risk?

This chapter is organized as follows. We introduce the issue of how to choose sample size for estimation of a parameter or for a hypothesis test regarding a parameter in the context of one-sample studies in which it is desired to estimate or test a population proportion. We continue on to two-sample studies involving comparisons between two proportions and one- and two-sample studies involving estimation or testing of population means. We conclude with a section on sample size for logistic regression.

In this chapter, we will provide a brief introduction to power and sample size computation and only address sample size issues for a few of the procedures that are most commonly used in epidemiological research. However, we do hope that the reader will gain a sense for what one can accomplish by planning a study with appropriate attention to sample size considerations.

A focus on sample size considerations when the study is first being planned is critical for the ultimate likelihood that a study proposal is accepted for funding and that the final manuscript will be accepted for publication. To ignore the issue of sample size would greatly increase the likelihood of embarking on a costly and time-consuming epidemiological study with little likelihood of finding any definitive results.

28.2 One-Group Designs, Inferences About Proportions

The simplest study design is one in which interest focuses on results for a single group. One is often interested in making inferences about the value of a population proportion. In this section, we will illustrate how to choose sample size for the following examples:

Example 28.1.
A district medical officer seeks to estimate the proportion of children in the district receiving appropriate childhood vaccinations. Assuming a simple random sample is to be selected from a community, how many children must be studied if the resulting estimate is to fall within 10 percentage points of the true proportion with 95% confidence?

Example 28.2.
Consider the information given in Example 28.1, only this time we will determine the sample size necessary to estimate the proportion vaccinated in the population to within 10% (not 10 percentage points) of the true value.

Example 28.3.
During a virulent outbreak of neonatal tetanus, health workers wish to determine whether the rate is decreasing after a period during which it had risen to a level of 150 cases per 1,000 live births. What sample size is necessary to test the null hypothesis that the population proportion is 0.15 at the 0.05 level if it is desired to have a 90% probability of detecting a decrease to a rate of 100 per 1,000 if that were the true proportion?

The first two examples involve estimation and confidence intervals, while the third involves a statistical hypothesis test.

The usual model underlying testing or estimation of a population proportion assumes that the design involves a simple independent random sample from a population in which the probability of a "success" is constant. The distribution of the number of successes in a sample of size n with a true underlying proportion of successes denoted by π is given by the binomial distribution. However, formulas are simplified when power and sample size determinations are made on the basis of using the normal approximation to the binomial.

The sampling distribution of the sample proportion "p" is approximately normal with mean of π (the expected value of p, $E\,(p) = \pi$) and variance of p, $\mathrm{Var}\,(p) = \pi\,(1-\pi)/n$; the standard deviation is $\sqrt{\pi\,(1-\pi)/n}$.

We begin by discussing sample size determination for estimation (the confidence interval approach) and then turn to sample size determination for hypothesis testing problems.

28.2.1 Confidence Intervals for a Single Population Proportion

Two-sided $100\,(1-\alpha)\,\%$ confidence intervals for a parameter, θ, based on using the normal approximation can be stated in general as

$$\hat{\theta} \pm z_{1-\alpha/2}\widehat{\mathrm{SE}}\left(\hat{\theta}\right), \tag{28.1}$$

where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the normal (or Gaussian) distribution. For the commonly used two-sided 95% confidence interval, $z_{1-\alpha/2} = 1.96$. The $100\,(1-\alpha)\,\%$ confidence interval for π based on the estimated proportion, p, is given by

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}. \tag{28.2}$$

Letting ω be the half-width of the confidence interval for the expected true value π, we have

$$\omega = z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}. \tag{28.3}$$

Table 28.1 Sample size for 95% two-sided confidence interval for a proportion (using the normal approximation) to have expected width, ω

	ω	
π	±0.05	±0.10
0.50	385	97
0.25	289	73
0.10	139	35

The sample size necessary to achieve a confidence interval of width ω is given by

$$n = \left(\frac{z_{1-\alpha/2}}{\omega}\right)^2 [\pi(1-\pi)]\,. \tag{28.4}$$

Returning to Example 28.1, we begin by assuming that the rate of vaccinated children is expected to be about 75%. We would then set $\pi = 0.75$, $\omega = 0.10$, and $z_{1-\alpha/2} = 1.96$. From Eq. 28.4, we find that $n = 72.03$. Note that for sample size calculations, we round up. We conclude that to estimate the expected population proportion to within ±0.10, a sample of 73 children would be required.

If we do not really know what rate to expect, we can make use of the fact that n will be largest for $\pi = 0.50$ and use this value to solve for n. For Example 28.1, we require a sample size of 97 to be sure that the confidence interval width will be no wider than plus or minus 10 percentage points no matter what the observed proportion is.

Table 28.1 presents the required sample sizes for selected values of π and ω.

Proceeding to Example 28.2, we consider the information given in Example 28.1, only this time we will determine the sample size necessary to estimate the proportion vaccinated in the population to within 10% (not 10 percentage points) of the true value.

Let θ be the unknown population parameter as before and let $\hat{\theta}$ be the estimate of θ. Let ε, the desired precision, be defined as

$$\varepsilon = \frac{\left|\hat{\theta} - \theta\right|}{\theta}.$$

In the present example, based on the confidence limits using the normal approximation to the distribution of p, it follows that

$$|p - \pi| = z_{1-\alpha/2}\frac{\sqrt{\pi\,(1-\pi)}}{\sqrt{n}}$$

and, dividing both sides by π, an expression similar to the one presented above for ε is obtained. That is,

$$\varepsilon = \frac{|p-\pi|}{\pi} = z_{1-\alpha/2}\frac{\sqrt{1-\pi}}{\sqrt{n\pi}},$$

and squaring both sides and solving for n gives

$$n = z_{1-\alpha/2}^{2} \frac{1-\pi}{\varepsilon^{2}\pi}. \tag{28.5}$$

Assuming $\pi = 0.75$, we would find that a sample size of 129 would be required to assure that the 95% confidence interval would be within 10% of the true value.

28.2.2 Hypothesis Testing for a Single Population Proportion

Suppose we would like to test a null hypothesis about the value of the population proportion

$$H_0 : \pi = \pi_0$$

versus the one-sided alternative hypothesis

$$H_a : \pi > \pi_0.$$

Statistical hypothesis testing involves balancing the two types of errors that can be made. Type I error is defined as the error of rejecting the null hypothesis when it is in fact true. We denote the probability of making a type I error as "α"; a commonly used choice for α is 0.05. The critical value of the test statistic is then chosen so that the probability of rejecting the null hypothesis when it is true will be α.

To choose the necessary sample size, we need to address type II error as well. A type II error is the error of failing to reject the null hypothesis when it is in fact false. To determine the probability of a type II error (denoted by "β"), we must specify a particular value of interest for the alternative hypothesis, say, π_a. The probability of rejecting the null hypothesis when it is false is defined as the *power* of the test, $1-\beta$. Typically, we require the power at the alternative of interest to be 80% or 90%.

Based on the normal approximation to the binomial, the test statistic for a test of the null hypothesis is given by

$$z = \frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}.$$

To set the probability of a type I error equal to α, we plan to reject the null hypothesis if $z > z_{1-\alpha}$. To choose n, we fix the probability that $z > z_{1-\alpha}$ if the population proportion equals π_a to be $1-\beta$. This may be represented graphically as shown in Fig. 28.1.

In this figure, the point "**c**" represents the upper 100αth percent point of the distribution of p for the sampling distribution centered at π_0 (i.e., the distribution which would result if the null hypothesis were true):

$$c = \pi_0 + z_{1-\alpha}\sqrt{\pi_0\,(1-\pi_0)/n}.$$

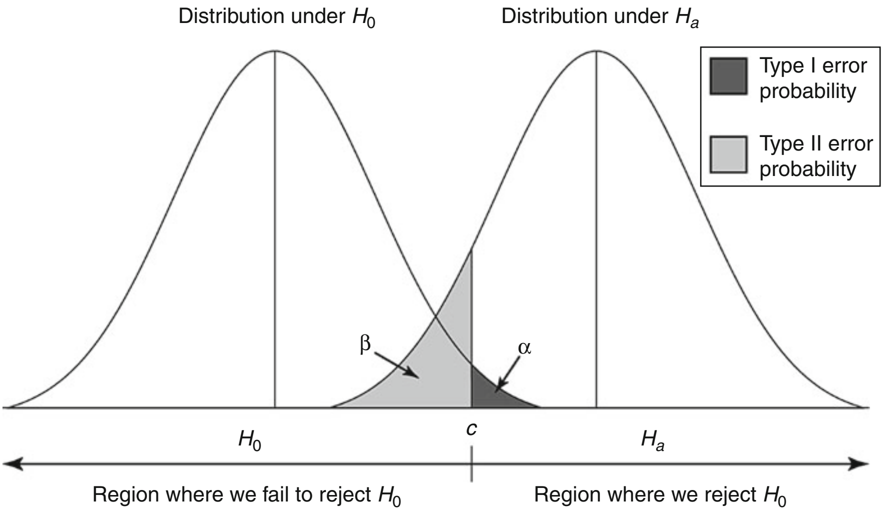

Distribution under H_0
Distribution under H_a
Type I error probability
Type II error probability
β
α
H_0
c
H_a
Region where we fail to reject H_0
Region where we reject H_0

Fig. 28.1 Sampling distributions for one-sample hypothesis test

For the sampling distribution centered at π_a (i.e., the distribution which would result if the alternate hypothesis were true), "**c**" represents the lower 100βth percent point of the distribution of p:

$$c = \pi_a + z_\beta \sqrt{\pi_a (1 - \pi_a)/n}.$$

In order to find n, we set the two expressions equal to each other. From this, it follows that

$$\pi_0 + z_{1-\alpha} \sqrt{\pi_0 (1 - \pi_0)/n} = \pi_a + z_\beta \sqrt{\pi_a (1 - \pi_a)/n}.$$

Noting that $z_{1-\beta} = -z_\beta$, we find

$$\pi_a - \pi_0 = \left\{ z_{1-\alpha} \sqrt{\pi_0 (1 - \pi_0)} + z_{1-\beta} \sqrt{\pi_a (1 - \pi_a)} \right\} \Big/ \sqrt{n},$$

and, solving for n, we find that the necessary sample size, for this single sample hypothesis testing situation, is given by the formula

$$n = \frac{\left\{ z_{1-\alpha} \sqrt{\pi_0 (1 - \pi_0)} + z_{1-\beta} \sqrt{\pi_a (1 - \pi_a)} \right\}^2}{(\pi_a - \pi_0)^2}. \tag{28.6}$$

Notice that as π_a gets further and further away from π_0, the necessary sample size decreases.

Table 28.2 Sample size for 0.05-level, two-sided test that the proportion equals π_0 versus the alternative π_a for specified levels of power (based on normal approximation)

		Power	
π_0	π_a	80%	90%
0.50	0.40	194	259
0.50	0.30	47	62
0.20	0.10	108	137
0.15	0.10	363	471
0.10	0.05	239	301

To illustrate, we return to Example 28.3 in which we wish to test the null hypothesis that $\pi = 0.15$ at the one-sided 5% level and have 90% power to detect a decrease to a rate of 0.10. Using Eq. 28.6, it follows that

$$n = \frac{\left\{ 1.645\sqrt{0.15\,(0.85)} + 1.282\sqrt{0.10\,(0.90)} \right\}^2}{(0.05)^2} = 377.90.$$

Hence, we see that a total sample size of 378 live births would be necessary.

To plan sample size for a two-sided test, we need only substitute $z_{1-\alpha/2}$ for $z_{1-\alpha}$ in Eq. 28.6 to obtain

$$n = \frac{\left\{ z_{1-\alpha/2}\sqrt{\pi_0\,(1-\pi_0)} + z_{1-\beta}\sqrt{\pi_a\,(1-\pi_a)} \right\}^2}{(\pi_a - \pi_0)^2}. \tag{28.7}$$

To have 90% power for a two-sided 5% level test for Example 28.3 would require a total of 471 subjects to detect the difference between the null hypothesis proportion, π_0, of 0.15 and the alternative proportion, π_a, of 0.10. Note that the sample size required to achieve 90% power for the specified alternative is larger when a two-sided 5% level test is planned than when a one-sided 5% level test is planned, so that the investigator needs to be clear as to whether the planned test is to be one-sided or two-sided when making sample size computations.

Table 28.2 presents the required sample sizes for selected values of π_0, π_a, and power. For a two-sided test, unless the null hypothesis proportion equals 0.5, computed sample sizes for alternative proportions given by $\pi_{aL} = \pi_0 - \delta$ and $\pi_{aU} = \pi_0 + \delta$ will differ; the larger estimate of sample size will be obtained for the alternative proportion closer to 0.5.

28.2.3 Additional Considerations and References

Good introductions to sample size computations for tests and confidence intervals for a single proportion can be found in Dixon and Massey (1983), Lemeshow et al. (1990), Fleiss (1981), and Lachin (1981). Books containing sample size tables are available (e.g., Machin and Campbell 1987; Machin et al. 1997; Lemeshow et al. 1990). Commercially available sample size software such as nQuery Advisor

Release 5 (Elashoff 2002) can be used to compute sample size for confidence intervals or hypothesis tests (based on either the normal approximation or an exact binomial test) for a single proportion as well as for a wide variety of other sample size problems.

For values of π near 0 or 1 (or for small sample sizes), sample size methods involving a continuity correction (Fleiss et al. 1980), methods designed for rare events (e.g., Korn 1986; Louis 1981), or methods based on exact tests (Chernick and Liu 2002) may be preferable.

Note that an actual field survey is unlikely to be based on a simple random sample. As a result, the required sample size would go up by the amount of the "design effect" which is determined by the details of the actual sampling plan. The "design effect" is the ratio of the standard error of the estimated parameter under the study design to the standard error of the estimate under simple random sampling; a text on sample surveys should be consulted for details (see Levy and Lemeshow 1999). For example, if a cluster sampling plan with a design effect of 2 were to be employed, the sample size computed using the above formulas would need to be doubled.

28.3 Comparison of Two Independent Proportions

28.3.1 Study Designs, Parameters, and Analysis Methods

More sample size literature exists for the problem of comparing two independent proportions than for any other sample size problem. This has come about because there are several basic sampling schemes leading to problems of this type. There are different parameterizations of interest and a variety of test and estimation procedures that have been developed. Sample size formulations depend on the parameter of interest for testing or estimation as well as the specifics of the test or estimation procedure.

The basic study designs relevant to epidemiological studies are experimental trials, cohort studies, and case-control studies. We describe each study type briefly and give an example. The examples will be addressed in more detail in subsequent sections.

Experimental Trial $2n$ subjects are recruited for a study; n are randomly assigned to group 1 and n to group 2. The intervention is applied according to the design. Subjects are followed for a fixed time and success-failure status is recorded. Experimental trials are usually randomized, often double blind, and always prospective. For example, patients with intestinal parasites are randomly assigned to receive either the standard drug or a new drug and followed to determine whether they respond favorably. The observed proportion responding favorably in group i is denoted by p_i and the true population proportion in group i by π_i.

Experimental trials are typically analyzed in terms of the difference in proportions, or the risk difference:

$$\text{Population risk difference} = \pi_1 - \pi_2. \tag{28.8}$$

$$\text{Estimated risk difference} = p_1 - p_2. \tag{28.9}$$

Cohort Study *n* subjects are recruited from group 1 and *n* from group 2; subjects are followed for a fixed time and success-failure status is recorded. Cohort studies are typically prospective studies. For example, workers with asbestos exposure and workers in the same industry without asbestos exposure are followed for the development of lung disease.

Cohort studies may be analyzed in terms of the risk difference or in terms of the relative risk:

$$\text{Population relative risk} = RR = \pi_2/\pi_1 \tag{28.10}$$

$$\text{Estimated relative risk} = rr = p_2/p_1. \tag{28.11}$$

Referring to the example, π_1 denotes the true proportion of diseased workers in the unexposed group, while π_2 denotes the true proportion of diseased workers in the exposed group, and p_1 and p_2 are the corresponding observed proportions.

Case-Control Studies *n* subjects (cases) are recruited from among those who have developed a disease, and *n* subjects (controls) are recruited from a similar group without the disease. Subjects from both groups are studied for the presence of a relevant exposure in their background. For example, tuberculosis (TB) cases and controls are assessed for whether they had been vaccinated with BCG (Bacillus Calmette-Guérin vaccine). Case-control studies are inherently retrospective studies, and interest is focused on the odds ratio:

$$\text{Population odds ratio} = OR = \pi_2\,(1 - \pi_1)/(1 - \pi_2)\pi_1. \tag{28.12}$$

$$\text{Estimated odds ratio} = or = p_2\,(1 - p_1)/(1 - p_2)p_1. \tag{28.13}$$

Referring to the example, π_1 denotes the true proportion of vaccinated subjects among the controls, while π_2 denotes the true proportion of vaccinated subjects among the TB cases, and p_1 and p_2 are the corresponding observed proportions.

We begin by discussing sample size determination for estimation (the confidence interval approach) and then turn to sample size determination for hypothesis testing problems.

28.3.2 Confidence Intervals for the Risk Difference

Example 28.4.

A pilot study with 20 subjects randomized to receive the standard drug to control intestinal parasites and 20 to receive a new drug found that 13 subjects (65%) receiving the standard drug responded favorably, while 17 (85%) of the subjects receiving the new drug responded favorably.

Question 4a: Do these data establish that the new drug is better (lower limit of confidence interval is greater than zero) and, if not, might it still be enough better to warrant a larger clinical trial? We address this question with a confidence interval below.

Question 4b: What sample size would be required for the larger clinical trial? We address this question in the context of a confidence interval later in this section and in the context of a hypothesis test in the following section.

The estimated value of the risk difference, $\pi_1 - \pi_2$, is given by $p_1 - p_2$, the observed difference in proportions. The variance of $p_1 - p_2$ for independent proportions when the sample sizes, n, in each group are equal is

$$\mathrm{Var}(p_1 - p_2) = \frac{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}{n}. \tag{28.14}$$

This formula is based on the assumption that the data come from independent random samples from the populations of interest. In population i, the probability of a success is a constant, π_i, and therefore the number of successes observed for each group has a binomial distribution with parameters n and π_i.

The standard error of this estimate, $p_1 - p_2$, is estimated by substituting the observed proportions for the true proportions and is given by

$$\mathrm{SE}(p_1 - p_2) = \frac{\sqrt{p_1(1-p_1) + p_2(1-p_2)}}{\sqrt{n}}. \tag{28.15}$$

Referring to the basic formula for a confidence interval based on the normal approximation given in Eq. 28.1, a two-sided 95% confidence interval for the difference in the proportions responding favorably to the new drug in comparison to the old drug is given by

$$0.85 - 0.65 \pm 1.96 \frac{\sqrt{0.85(1-0.85) + 0.65(1-0.65)}}{\sqrt{20}}.$$

The limits are 0.20 ± 0.209 or –0.009 to 0.409, suggesting that although we cannot rule out a difference of zero, the data indicate that the new drug might work markedly better than the standard.

The investigator wants to plan a definitive study to assess how much the success rates really do differ. What sample size would be necessary to obtain a confidence interval whose width is less than or equal to ±0.05?

We require that the confidence interval for $\pi_1 - \pi_2$, be $p_1 - p_2 \pm \omega$, where for Example 28.4, $\omega \le 0.05$. To obtain a confidence interval width satisfying these conditions, we must have

$$z_{1-\alpha/2} \frac{\sqrt{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}}{\sqrt{n}} \le \omega.$$

Table 28.3 Sample size per group for 95% two-sided confidence interval (using normal approximation) for risk difference to have expected width, ω

		ω	
π_1	π_2	± 0.05	± 0.10
0.50	0.50	769	193
0.50	0.25	673	169
0.50	0.10	523	131
0.25	0.25	577	145
0.25	0.10	427	107
0.10	0.10	277	70

Solving this equation for n, the sample size in each group, we obtain Eq. 28.16:

$$n = \frac{z_{1-\alpha/2}^2\left[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)\right]}{\omega^2}. \tag{28.16}$$

For Example 28.4, an n per group of 546 would be required to obtain an expected 95% two-sided confidence interval width of approximately ±0.05 if we expect to see about the same proportions as we did in the pilot study.

Table 28.3 presents the sample size in each group necessary to obtain specified confidence interval widths for a few selected examples. This table should provide investigators with a quick idea of the order of magnitude of required sample sizes. Note that since the confidence interval width depends on the postulated proportions only through the terms $\pi_i\ (1-\pi_i)$, this table can also be used for proportions greater than 0.5.

If an investigator is a bit uncertain about what proportions to expect and wants to ensure that the confidence interval width is less than some specified amount $\pm\omega$ no matter what proportions are observed, we can use the fact that the confidence interval is widest when $\pi_1 = \pi_2 = 0.5$. In this case, the sample size required for each group is

$$n \le \frac{z_{1-\alpha/2}^2}{2\,\omega^2}. \tag{28.17}$$

For a two-sided 95% confidence interval, this becomes approximately $2/\omega^2$. For Example 28.4, the maximum sample size per group required for a confidence interval width of no more than ±0.05 is 769.

28.3.3 Confidence Interval for Relative Risk (Ratio)

Example 28.5.
Workers with asbestos exposure and workers in the same industry without asbestos exposure are followed for the development of lung disease. Suppose that disease occurs in 20% of the unexposed group, how large a sample would be needed in each of the exposed and unexposed study groups to estimate the relative risk to within 10% of the true value with 95% confidence assuming that the relative risk is approximately 1.75?

Forthis purpose, we define group 1 as the unexposed group and group 2 as the exposed group. The estimate of the relative risk (cf. chapter ▶Rates, Risks, Measures of Association and Impact of this handbook) is

$$\widehat{RR} = rr = p_2/p_1.$$

Since we are dealing with a ratio, which can be expected to have a skewed distribution with a lognormal shape, we need to take logs to normalize the distribution so that the normal approximation can be used to construct the confidence interval.

We obtain the standard deviation for the estimate for the case where the sample sizes in the two groups are equal by using the approximation

$$\mathrm{Var}(\ln(rr)) \quad \approx \quad \frac{1-\pi_1}{n\pi_1} + \frac{1-\pi_2}{n\pi_2}. \tag{28.18}$$

The estimated standard deviation is obtained by substituting the estimated proportions for the population proportions and taking the square root.

The 100 $(1-\alpha)$ % confidence limits for ln (RR) are given by ln $(rr) \pm \omega$ where

$$\omega = z_{1-\alpha/2}\widehat{SE}\left(\ln(rr)\right) \quad = \quad z_{1-\alpha/2}\sqrt{\frac{1-\pi_1}{n\pi_1} + \frac{1-\pi_2}{n\pi_2}}.$$

Then the confidence limits for RR are given by exp $(\ln(rr_L))$ and exp $(\ln(rr_U))$, where ln (rr_L) and ln (rr_U) are the lower and upper confidence limits for ln (RR).

To choose the sample size necessary to obtain a confidence interval of a desired width for ln (RR), we could simply specify ω and solve for n:

$$n = \frac{z^2{}_{1-\alpha/2}\left[\frac{1-\pi_1}{\pi_1} + \frac{1-\pi_2}{\pi_2}\right]}{\omega^2}. \tag{28.19}$$

Alternatively, an investigator may wish to specify the width in terms of how close the limits are to RR. For example, suppose that we are thinking in terms of values of $RR > 1$ and that we want the difference between RR and RR_L to be no greater than εRR; that is, we set $RR - RR_L = \varepsilon RR$ which we rearrange to get $RR\,(1-\varepsilon) = RR_L$. Then, taking logs, we have $\ln(RR) + \ln(1-\varepsilon) = \ln(RR_L)$ and $\ln(RR) - \ln(RR_L) = -\ln(1-\varepsilon) = \omega$, so

$$\omega = z_{1-\alpha/2}\sqrt{\frac{1-\pi_1}{n\,\pi_1} + \frac{1-\pi_2}{n\,\pi_2}} = -\ln(1-\varepsilon)\,.$$

Then to find the necessary sample size for each group, we solve for n to obtain

$$n = \frac{z^2{}_{1-\alpha/2}\left[\frac{1-\pi_1}{\pi_1} + \frac{1-\pi_2}{\pi_2}\right]}{[\ln(1-\varepsilon)\,]^2}. \tag{28.20}$$

Table 28.4 Sample size per group for 95% two-sided confidence interval for the relative risk to have lower limit $(1-\varepsilon)RR$

		ε	
RR	π_1	0.10	0.20
1.25	0.20	2,423	540
1.50	0.20	2,192	489
1.75	0.20	2,027	452
2.00	0.20	1,904	424
1.25	0.40	866	193

A version of this, which substitutes the expected RR for π_2, is

$$n = \frac{z^2{}_{1-\alpha/2}\left[\frac{1+RR}{RR\pi_1} - 2\right]}{[\ln(1-\varepsilon)]^2}. \tag{28.21}$$

Returning to Example 28.5, the expected $RR = 1.75$, and $\pi_1 = 0.20$, and we have requested that the lower limit of the confidence interval for RR be within 10% of the true value of RR. Therefore, $\varepsilon = 0.1$, and $1-\varepsilon = 0.9$, and the required sample size would be 2,027 per group or 4,054 total.

Table 28.4 presents the sample size in each group necessary to obtain specified confidence interval widths for a few selected examples.

28.3.4 Confidence Intervals for the Odds Ratio

Example 28.6.
The efficacy of BCG vaccine in preventing childhood tuberculosis is in doubt, and a study is designed to compare the immunization coverage rates in a group of tuberculosis cases compared to a group of controls. Available information indicated that roughly 30% of controls are not vaccinated, and we wish to estimate the odds ratio to within 20% of the true value. It is believed that the odds ratio is likely to be about 2.0.

For problems involving estimation of the odds ratio (cf. chapter ▶Rates, Risks, Measures of Association and Impact of this handbook), we let group 1 denote the controls and group 2 denote the cases. Our estimate of the odds ratio is

$$or = \frac{p_2(1-p_1)}{(1-p_2)\,p_1}.$$

Since we are dealing with a ratio, we need to take logs so that the normal approximation can be used to construct the confidence interval.

We obtain the standard deviation for the estimate for the case where the sample sizes in the two groups are equal, by using the approximation

$$\mathrm{Var}(\ln(or)) \approx \frac{1}{n\pi_1(1-\pi_1)} + \frac{1}{n\pi_2(1-\pi_2)}. \tag{28.22}$$

Table 28.5 Sample size per group for 95% two-sided confidence interval for *OR* to have lower limit $(1-\varepsilon)\,OR$

		ε	
OR	π_1	0.10	0.20
1.25	0.30	3,171	708
1.50	0.30	3,101	692
1.75	0.30	3,061	683
2.00	0.30	3,040	678
1.25	0.50	2,786	621

The estimated standard deviation is obtained by substituting the estimated proportions for the population proportions and taking the square root.

To obtain a $100\,(1-\alpha)\,\%$ confidence interval for $\ln(OR)$ of width ω where $\omega = z_{1-\alpha/2}\mathrm{SE}\,(\ln(or))$ when the sample sizes in the two groups are equal, we require a sample size per group of

$$n = \frac{z^2_{1-\alpha/2}\left[\frac{1}{\pi_2(1-\pi_2)} + \frac{1}{\pi_1(1-\pi_1)}\right]}{\omega^2}. \tag{28.23}$$

In situations where we assume that the odds ratio is greater than 1.0, to specify that the lower limit of the confidence interval be no less than $(1-\varepsilon)\,OR$, we would set $\omega = -\ln(1-\varepsilon)$ as we did in the previous section for the relative risk. We then obtain

$$n = \frac{z^2_{1-\alpha/2}\left[\frac{1}{\pi_2(1-\pi_2)} + \frac{1}{\pi_1(1-\pi_1)}\right]}{[\ln(1-\varepsilon)]^2}. \tag{28.24}$$

Solving for π_2 using Eq. 28.12, we have

$$\pi_2 = \frac{OR\,\pi_1}{OR\,\pi_1 + (1-\pi_1)}$$

and we can obtain sample size expressed in terms of π_1 and OR:

$$n = z^2_{1-\alpha/2}\left[\frac{OR + (1-\pi_1+OR\pi_1)^2}{\pi_1(1-\pi_1)OR[\ln(1-\varepsilon)]^2}\right]. \tag{28.25}$$

For Example 28.6, we have $OR = 2$, $\pi_1 = 0.30$, $(\pi_2 = 0.462)$, and $(1-\varepsilon) = 0.8$, so we need 678 subjects per group.

Table 28.5 presents the sample size in each group necessary to obtain specified confidence interval widths for *OR* for a few selected examples.

28.3.5 Testing the Difference Between Two Proportions

The goal is to test

$H_0 : \pi_1 = \pi_2$ versus $H_1 : \pi_1 \neq \pi_2$.

If it can be assumed that the samples of size n from both groups arise from independent binomial distributions, the test for H_0 can be performed using the normal approximation to the binomial.

The test statistic is

$$z = \frac{\sqrt{n}\,(p_1 - p_2)}{\sqrt{2\bar{p}\,(1-\bar{p})}}, \tag{28.26}$$

where $z \sim N\,(0, 1)$, i.e., z is normally distributed with mean 0 and variance 1, and where, in the general case with unequal sample sizes in the two groups,

$$\bar{p} \;=\; \frac{n_1 p_1 \;+\; n_2 p_2}{n_1 \;+ n_2},$$

whereas for equal sample sizes,

$$\bar{p} \;=\; \frac{p_1 \;+\; p_2}{2}.$$

(Note that the two-sided z test given by Eq. 28.26 is algebraically equivalent to the standard χ^2 test.)

The sample size in each group required for a two-sided $100\,(1-\alpha)\,\%$ test to have power $1-\beta$ is

$$n = \frac{\left[z_{1-\alpha/2}\,\sqrt{2\overline{\pi}(1-\overline{\pi})} + z_{1-\beta}\,\sqrt{\pi_1\,(1-\pi_1) + \pi_2\,(1-\pi_2)}\right]^2}{(\pi_1 - \pi_2\,)^2}. \tag{28.27}$$

and $\overline{\pi}$ is defined by analogy with $\overline{p}$.

Example 28.7.
Typically, the outcome measure for placebo controlled double-blind trials for acute duodenal ulcer healing is the proportion of patients whose ulcer has healed by 4 weeks as ascertained by endoscopy. The healing rate for the placebo group is typically about 40%. H_2-blocking active drugs usually result in 70% healed. The investigator wishes to evaluate a new drug with the expectation of seeking FDA (US Food and Drug Administration) approval; the results will be assessed by comparing observed proportions healed using the χ^2 test at the two-sided 5% significance level. Such trials are expensive to mount so that if the new drug is as effective as those currently approved, the investigator wants a 90% chance that the trial will yield a significant result.

Using Eq. 28.27 for a two-sided 5% test, a sample size of 56 patients per group or a total sample size of 112 patients would be required to achieve 90% power.

Table 28.6 presents the sample size in each group necessary for a 5% two-sided χ^2 test comparing two independent proportions to have specified power for a few selected examples.

Table 28.6 Sample size per group for 5% two-sided χ^2 test for the difference between two independent proportions to have specified power

		Power	
π_1	π_2	80%	90%
0.10	0.05	435	582
0.25	0.10	100	133
0.50	0.25	58	77
0.50	0.10	20	26

Table 28.7 Sample size per group for 5% two-sided test that the relative risk equals 1 to have specified power

		Power	
RR	π_1	80%	90%
1.25	0.20	1,094	1,464
1.50	0.20	294	392
1.75	0.20	138	185
2.00	0.20	82	109
1.25	0.40	388	519

28.3.6 Testing the Relative Risk

In a cohort study, where we want to focus attention on a test of the relative risk

$$H_0 : RR = \frac{\pi_2}{\pi_1} = 1,$$

the large sample test for this null hypothesis is the same as for the null hypothesis that the difference in proportions is zero and therefore the sample size formulas are the same. If we substitute *RR* into Eq. 28.27, we obtain

$$n = \frac{\left[z_{1-\alpha/2}\sqrt{(1+RR)[1-\pi_1(1+RR)/2]} + z_{1-\beta}\sqrt{[1+RR-\pi_1(1+RR^2)]}\right]^2}{\pi_1(1-RR)^2}. \tag{28.28}$$

Example 28.8.
Two competing therapies for a particular cancer are to be evaluated by the cohort study strategy in a multicenter clinical trial. Patients are randomized to either treatment A or B and are followed for recurrence of disease for 5 years following treatment. How many patients should be studied in each of the two arms of the trial in order to have 90% power to reject $H_0 : RR = 1$ in favor of the alternative $RR = 0.5$, if the test is to be performed at the two-sided $\alpha = 0.05$ level and it is assumed that $\pi_1 = 0.35$?

For Example 28.8, we substitute $\pi_1 = 0.35$ and $RR = 0.5$ into Eq. 28.28 and find that the required sample size per group would be 131 or 262 total. Or we could have noted that $\pi_2 = 0.175$ and used Eq. 28.27.

Table 28.7 presents the sample size in each group necessary for a 5% two-sided normal approximation test of the null hypothesis that the relative risk is 1.0 to have specified power for a few selected examples.

28.3.7 Testing the Odds Ratio

The null hypothesis that the odds ratio equals 1.0 can be tested using Eq. 28.26 as for the test of difference in proportions. Sample size formulas can be modified to be based on π_2 and OR by algebraic substitution in Eq. 28.27 if desired; however, formulas are simpler if we use Eq. 28.12 to solve for the other proportion and use Eq. 28.27 directly.

Example 28.9.
The efficacy of BCG vaccine in preventing childhood tuberculosis is in doubt, and a study is designed to compare the immunization coverage rates in a group of tuberculosis cases compared to a group of controls. Available information indicates that roughly 30% of the controls are not vaccinated, and we wish to have an 80% chance of detecting whether the odds ratio is significantly different from 1 at the 5% level. If an odds ratio of 2 would be considered an important difference between the two groups, how large a sample should be included in each study group?

For Example 28.9, $\pi_1 = 0.3$ and $OR = 2$ and thus $\pi_2 = 0.462$, so using Eq. 28.27, we find that to obtain 80% power for a two-sided 5% level test would require 141 subjects per group or 282 total.

Table 28.8 presents the sample size in each group necessary to obtain specified power for tests of $OR = 1$ for a few selected examples.

28.3.8 Additional Considerations and References

Good introductions to sample size computations for tests and confidence intervals for comparing two independent proportions can be found in Dixon and Massey (1983), Lemeshow et al. (1990), Fleiss (1981), and Lachin (1981). Books containing sample size tables are available (e.g., Machin and Campbell 1987; Machin et al. 1997; Lemeshow et al. 1990). Commercially available sample size software such as nQuery Advisor Release 5 (Elashoff 2002) can be used to compute sample size (or width) for confidence intervals and sample size (or power) for hypothesis tests for the two proportion case (based on either the normal approximation, continuity corrected normal approximation, or Fisher's exact test) as well as for a wide variety of other sample size problems.

For values of π near 0 or 1 (or for small sample sizes), sample size methods involving a continuity correction (Fleiss et al. 1980), or methods based on exact tests (Chernick and Liu 2002) may be preferable.

Table 28.8 Sample size per group for 5% two-sided test of $OR = 1$ to have specified power

		Power	
OR	π_1	80%	90%
1.25	0.30	1,442	1,930
1.50	0.30	425	569
1.75	0.30	219	293
2.00	0.30	141	188
1.25	0.50	1,267	1,695

When plans call for the sample sizes in the two groups to be unequal, the formulas for sample size and power must incorporate the expected ratio of the sample sizes, see references above. Generally, for the same total sample size, power will tend to be higher and confidence interval widths narrower when sample sizes are equal; for comparisons of proportions, total sample size will depend on whether the proportion closer to 0.5 has the larger or the smaller sample size.

Note that the sample size methods discussed above do not apply to correlation/agreement/repeated measures (or pair-matched case-control) studies in which N subjects are recruited and each subject is measured by two different raters, or is studied under two different treatments in a crossover design. These designs cannot be analyzed using the methods described for independent proportions; for example, sample size computations for the difference between two correlated proportions are based on the McNemar test (Lachin 1992).

28.4 One-Group Designs, Inferences About a Single Mean

We turn to consideration of continuous outcomes and to inferences about the population mean. We denote the true but unknown mean in the population by μ and assume that the standard deviation for the population is given by σ. For a random sample of size n from a population with a normal (Gaussian) distribution, the distribution of the observed sample mean, $\bar{x}$, will also be normal with mean μ and standard deviation (also referred to as the standard error) given by $\mathrm{SE}(\bar{x}) = \sigma \big/ \sqrt{n}$. By the central limit theorem, the sampling distribution of the sample mean can usually be expected to be approximately normal for sample sizes of 30 or above even when the underlying population distribution is not normal.

28.4.1 Confidence Intervals for a Single Mean

Example 28.10.
Suppose an estimate is desired of the average retail price of 20 tablets of a commonly used tranquilizer. A random sample of retail pharmacies is to be selected. The estimate is required to be within 10 cents of the true average price with 95% confidence. Based on a small pilot study, the standard deviation in price, σ, can be estimated as 85 cents. How many pharmacies should be randomly selected?

Using the normal approximation, the two-sided $100\,(1-\alpha)\,\%$ confidence interval for the true mean, μ, for the case where the standard deviation is known, is given by

$$\bar{x} \;\pm\; z_{1-\alpha/2}\,\sigma \,/\sqrt{n}. \tag{28.29}$$

So the sample size required to obtain a confidence interval of width ω is

$$n \;=\; \frac{z^2_{1-\alpha/2}\,\sigma^2}{\omega^2}. \tag{28.30}$$

For Example 28.10, expressing costs in dollars,

$$n = \frac{(1.96)^2\,(0.85)^2}{(0.10)^2} = 277.6.$$

Therefore, a sample size of 278 pharmacies should be selected.

We should note however that usually the standard deviation must be estimated from the sample. Then the actual confidence interval for a sample mean would be given by

$$\bar{x} \pm t_{n-1,\,1-\alpha/2}\, s\,/\sqrt{n}, \tag{28.31}$$

where s is the observed standard deviation and $t_{n-1,1-\alpha/2}$ denotes the $100\,(1-\alpha/2)$th percentile of the t distribution with $n-1$ degrees of freedom. The value of $t_{n-1,1-\alpha/2}$ is always greater than $z_{1-\alpha/2}$; the values are close for large n, but t may be considerably larger than z for very small samples.

The required sample size would need to be larger than given by Eq. 28.30 simply to reflect the fact that $t_{n-1,1-\alpha/2} > z_{1-\alpha/2}$. In addition, the value of the standard deviation estimated from the sample will differ from the true standard deviation. The observed value of s may be either smaller or larger than the true value of the standard deviation, σ, and it can be expected to be larger than σ in about half of samples. So, even for large n, the observed confidence interval width will be greater than the specified ω in about half of planned studies.

To ensure that the observed confidence width will be shorter than ω more than half the time, we must take the distribution of s into account in the sample size computations. To solve for the required sample size for a confidence interval whose width has a specified probability, $1-\gamma$, of being narrower than ω requires the use of sample size software since an iterative solution based on the F and χ^2 distributions must be used (Kupper and Hafner 1989).

Returning to Example 28.10, specifying, for example, in nQuery Advisor, that the observed confidence interval width needs to be shorter than 0.1 with a probability of 50% ($1-\gamma = 0.5$) yields a required sample size of 280, only slightly larger than that given by Eq. 28.30. However, to increase the likelihood that the observed confidence interval width will be shorter than ω from 50% to 90% would require an increase in sample size from 280 to 309 (see Table 28.9).

Table 28.10 shows the required sample sizes for two-sided 95% confidence intervals to have specified widths (expressed in terms of ω/σ).

Table 28.9 Confidence interval for mean based on t (with coverage probability)

	1	2
Confidence level, $1-\alpha$	0.950	0.950
1- or 2-sided interval?	2	2
Coverage probability, $1-\gamma$	0.500	0.900
Standard deviation, σ	0.850	0.850
Distance from mean to limit, ω	0.100	0.100
N	280	309

Table 28.10 Sample size for 95% two-sided confidence interval for a single mean to have width less than or equal to ω with probability

	$100\,(1-\gamma)$	
ω/σ	50%	90%
0.05	1,539	1,609
0.10	386	421
0.20	98	116
0.30	45	56
0.50	18	24

Note that nQuery Advisor has been used to compute the sample sizes displayed in all the rest of the tables in this chapter.

28.4.2 Hypothesis Testing for a Single Population Mean

Suppose we would like to test the hypothesis

$$H_0 : \mu = \mu_0$$

versus the alternative hypothesis

$$H_a : \mu > \mu_0$$

and we would like to fix the level of the type I error to equal α and the type II error to equal β. That is, we want the power of the test to equal $1 - \beta$. We denote the actual value of the population mean under the alternative hypothesis as μ_a. Following the same development as for hypothesis testing about the population proportion (with the additional assumption that the variance of $\bar{x}$ is equal to σ^2/n under both H_0 and H_a), the necessary sample size for this hypothesis testing situation is given by

$$n = \frac{\sigma^2 \left[z_{1-\alpha} + z_{1-\beta}\right]^2}{\left[\mu_0 - \mu_a\right]^2}. \tag{28.32}$$

Alternatively, defining the effect size as

$$\delta = \frac{\mu_0 - \mu_a}{\sigma}, \tag{28.33}$$

we have

$$n = \frac{\left[z_{1-\alpha} + z_{1-\beta}\right]^2}{\delta^2}. \tag{28.34}$$

Example 28.11.
Pre- and poststudies with placebo in a variety of studies indicated that the standard deviation of blood pressure change was about 6 mmHg and that the mean reduction in the placebo group was typically close to 5 mmHg. To make a preliminary estimate of the value of a new intervention designed to lower blood pressure, it was planned to enroll subjects and

test the null hypothesis that mean reduction was 5 mmHg. The new intervention would be of interest if the mean reduction was 10 or greater. How large a sample would be necessary to test, at the 5% level of significance with a power of 90%, whether the average blood pressure reduction is 5 mmHg versus the alternative that the reduction is 10 mmHg when it is assumed that the standard deviation is 6 mmHg?

Using Eq. 28.32, we have

$$n = \frac{6^2\,(1.645 + 1.282)^2}{(10 - 5)^2} = 12.33.$$

Therefore, a sample of 13 patients with high blood pressure would be required.

A similar approach is followed when the alternative is two-sided. That is, when we wish to test

$$H_0 : \mu = \mu_0$$

versus

$$H_a : \mu \neq \mu_0.$$

In this situation, the null hypothesis is rejected if $\bar{x}$ is too large or too small. We assign area $\alpha/2$ to each tail of the sampling distribution under H_0. The only adjustment to Eq. 28.32 is that $z_{1-\alpha/2}$ is used in place of $z_{1-\alpha}$ resulting in

$$n = \frac{\sigma^2 \left[z_{1-\alpha/2} + z_{1-\beta}\right]^2}{\left[\mu_0 - \mu_a\right]^2}. \tag{28.35}$$

Returning to Example 28.11, a two-sided test could be used to test the hypothesis that the average reduction in blood pressure is 5 mmHg versus the alternative that the average reduction in blood pressure has increased and that a reduction of 10 mmHg would be considered important. Using Eq. 28.35 with $z_{1-\alpha/2} = 1.960$, $z_{1-\beta} = 1.282$, and $\sigma = 6$,

$$n = \frac{6^2\,(1.960 + 1.282)^2}{(10 - 5)^2} = 15.1.$$

Thus, 16 patients would be required for the sample if the alternative were two-sided.

Since usually the true standard deviation is unknown, a more accurate solution for the necessary sample size would require use of sample size software (computations are based on the central and non-central t distributions). Unlike the situation for confidence intervals, the normal approximation formula works well for computing sample size for a test; its accuracy can be improved by adding the correction factor

$$\frac{z^2_{1-\alpha/2}}{2} \tag{28.36}$$

before rounding up. For Example 28.11, this would lead to a sample size estimate of 18 (which agrees with the result given by nQuery Advisor).

Table 28.11 Sample size for two-sided 5% level *t* test to detect effect size $\delta = \mu_1 - \mu_2/\sigma$

δ	80% Power	90% Power
0.2	199	265
0.4	52	68
0.6	24	32
0.8	15	19
1.0	10	13
1.2	8	10

Table 28.11 presents the sample sizes necessary for 80% or 90% power for two-sided 5% level tests for specified effect sizes, δ.

28.5 Comparison of Two Independent Means

28.5.1 Confidence Intervals for the Difference Between Two Means

The difference between two population means is represented by a new parameter, $\mu_1 - \mu_2$. An estimate of this parameter is given by the difference in the sample means, $\bar{x}_1 - \bar{x}_2$. The mean of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is

$$E\left(\bar{x}_1 - \bar{x}_2\right) = \mu_1 - \mu_2,$$

and the variance of this distribution when the two samples are independent is

$$\mathrm{Var}\left(\bar{x}_1 - \bar{x}_2\right) = \mathrm{Var}\left(\bar{x}_1\right) + \mathrm{Var}\left(\bar{x}_2\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

where n_1 and n_2 are the sample sizes in the two groups.

In order for the distribution of the difference in sample means, $\bar{x}_1 - \bar{x}_2$, to have a *t* distribution, we must assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$. When the variances are equal and both sample sizes are equal to n, the formula for the variance of the difference can be simplified to

$$\mathrm{Var}\left(\bar{x}_1 - \bar{x}_2\right) = \frac{2\sigma^2}{n}.$$

The value σ^2 is an unknown population parameter, which can be estimated from sample data by pooling the individual sample variances, s_1^2 and s_2^2, to form the **pooled variance**, s_p^2, where, in the general case,

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 - 1) + (n_2 - 1)}.$$

Example 28.12.
Nutritionists wish to estimate the difference in caloric intake at lunch between children in a school offering a hot school lunch program and children in a school that does not. From

other nutrition studies, they estimate that the standard deviation in caloric intake among elementary school children is 75 calories and they wish to make their estimate to within 20 calories of the true difference with 95% confidence.

Using the normal approximation, the two-sided $100\,(1-\alpha/2)\,\%$ confidence interval for the true mean, $\mu_1 - \mu_2$, is given by

$$\bar{x}_1 \; - \; \bar{x}_2 \pm \; z_{1-\alpha/2}\, 2\sigma \,/\sqrt{n}. \tag{28.37}$$

So the sample size in each group required to obtain a confidence interval of width ω is

$$n \;=\; \frac{z^2_{1-\alpha/2}\, 2\sigma^2}{\omega^2}. \tag{28.38}$$

For Example 28.12,

$$n \;=\; \frac{(1.96)^2\,(2)(75)^2}{(20)^2} \;=\; 108.05.$$

Thus, a sample size of 109 children from each school should be selected.

We note, however, that the actual confidence interval for the difference in sample means would be given by

$$\bar{x}_1 \; - \; \bar{x}_2 \pm \; t_{2n-2,\,1-\alpha/2}\, s_p \sqrt{2}\,/\sqrt{n}, \tag{28.39}$$

where s_p is the observed pooled standard deviation and $t_{2n-2,1-\alpha/2}$ denotes the $100\,(1-\alpha/2)$ percentile of the t distribution with 2(n-1) degrees of freedom. So, as explained in the section on confidence intervals for a single mean, to solve for the required sample size for a confidence interval whose width has a specified probability, $1-\gamma$, of being narrower than ω requires the use of sample size software.

For Example 28.12, we show in Table 28.12 (pasted from nQuery Advisor) that a sample of 109 per group provides a 50% probability that the observed 95% confidence interval will have half-width less than 20, while to have a 90% probability that the confidence interval half-width will be less than 20 would require a sample of 123 children per school.

Table 28.12 Confidence interval for difference of two means (coverage probability) (equal n's)

	1	2
Confidence level, $1-\alpha$	0.950	0.950
1- or 2-sided interval?	2	2
Coverage probability, $1-\gamma$	0.500	0.900
Common standard deviation, σ	75.000	75.000
Distance from difference to limit, ω	20.000	20.000
n per group	109	123

Table 28.13 Sample size per group for 95% two-sided confidence interval for the difference in means to have width less than or equal to $\pm\omega$ with probability $(1-\gamma)$

	$100\,(1-\gamma)$	
ω/σ	50%	90%
0.05	3,075	3,145
0.10	770	805
0.20	193	211
0.30	87	98
0.50	36	39

Table 28.13 presents the sample sizes in each group required so that the two-sided 95% confidence interval for the difference in two independent means will be no wider than $\pm\omega$ with probability $(1-\gamma)$.

28.5.2 Testing the Difference Between Two Means (Two-Sample *t* Test)

The two-sample t test is used to test hypotheses about the population means in two independent groups of subjects. It is based on the assumptions that the underlying population distributions have equal standard deviations and that the population distributions are Gaussian (normal) in shape or that the sample sizes in each group are large. (In most cases, the distribution of the sample mean will be approximately Gaussian for sample sizes greater than 30.)

We consider tests of the null hypothesis:

$$H_0 : \mu_1 = \mu_2 \text{ or}$$

$$H_0 : \mu_1 - \mu_2 = 0$$

versus either

$$H_a : \mu_1 \neq \mu_2 \text{ for a two-sided test or}$$

$$H_a' : \mu_1 > \mu_2 \text{ or } H_a'' : \mu_1 < \mu_2 \text{ for one-sided tests.}$$

To avoid repetitions of formulas with minor changes, we write formulas only in terms of a two-sided test.

The sample size required in each group, to achieve a power of $1-\beta$, is

$$n = \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{(\mu_1 - \mu_2)^2}. \tag{28.40}$$

Setting

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}, \tag{28.41}$$

where δ is the effect size, we have a simpler version

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2}. \tag{28.42}$$

Table 28.14 Sample size in each group for two-sided 5% level t test to have specified power

δ	80% Power	90% Power
0.2	394	527
0.4	100	133
0.6	45	60
0.8	26	34
1.0	17	23
1.2	12	16

To improve the approximation, the correction factor in Eq. 28.43 may be added to Eq. 28.42 before rounding up:

$$\frac{z^2_{1-\alpha/2}}{4}. \tag{28.43}$$

Example 28.13.
A two-group, randomized, parallel, double-blind study is planned in elderly females after hip fracture. Patients will be studied for 2 weeks; each patient will be randomly assigned to receive either new drug or placebo three times per week. The sample sizes in the two groups will be equal. Plans call for a 5% level two-sided t test. The outcome variable will be change in hematocrit level during the study. Prior pilot data from several studies suggest that the standard deviation for change will be about 2.0% and it would be of interest to detect a difference of 2.2% in the changes observed in placebo and treated groups. What sample size in each group would be required to achieve a power of 90%?

For Example 28.13, the effect size is $2.2/2 = 1.1$. Using Eq. 28.42, we find

$$n = \frac{2(1.96 + 1.28)^2}{(1.1)^2} = 17.4.$$

Adding the correction factor of 0.96 and rounding up, we have a required sample size of 19 per group, which is the solution given using nQuery Advisor (computations are based on iterative methods and the central and non-central t, see Dixon and Massey 1983 or O'Brien and Muller 1983).

Table 28.14 shows the sample size needed in each group for a two-sided 5% level t test to achieve 80% or 90% power for the specified alternative, δ.

28.5.3 Additional Considerations and References

Good introductions to sample size computations for tests and confidence intervals for a single mean or for comparing two independent means can be found in Dixon and Massey (1983), O'Brien and Muller (1983), Lemeshow et al. (1990), Lachin (1981), and Rosner (2000). Books containing sample size tables are available (e.g., Machin and Campbell 1987; Machin et al. 1997). Commercially available sample size software such as nQuery Advisor Release 5 (Elashoff 2002) can be used to compute sample size (or width) for confidence intervals and sample size or power for hypothesis tests for means for either the single-group or two-group designs, as well as for a wide variety of other sample size problems.

When plans call for the sample sizes in the two groups to be unequal, the formulas for sample size and power must incorporate the expected ratio of the sample sizes, see references above. For the two-sample t test, for any given total sample size, N, power will be highest when both groups have the same sample size. For this reason, we generally prefer to plan equal sample sizes for a two-group study. However, sometimes investigators wish to plan a study with unequal n's; perhaps one type of subject is easier to accrue, or perhaps the investigator wants to maximize the number of subjects receiving the presumably superior treatment, or to accumulate extra safety information for the new treatment.

When the standard deviations in the two groups are markedly unequal, the usual t test with pooled variances is no longer the appropriate test. In many situations, the standard deviations show a patterned lack of homogeneity in which groups with higher means have higher standard deviations. In such cases, it is frequently advisable that sample size predictions (and later analysis) should be done on a transformed version of the variable. If the relationship between variance and mean is linear, this suggests using the square root of the variable. Such a transformation is likely to be desirable if the data represent counts or areas (note that the variable cannot be less than zero). If the relationship between standard deviation and mean is linear, this suggests using the log of the variable. This transformation is likely to be desirable for biological measures like viral load, triglyceride level, or variables ranging over several orders of magnitude (note that the variable cannot be negative or zero). If transformation does not seem to provide a solution to the problem of inequality of variances, it is possible that comparison of means is no longer the most appropriate method of analysis to address the question of interest. Assuming that transformation is not useful and comparison of means using a two-sample t test is still deemed appropriate, a modification of the t test may be planned; see, for example, Moser et al. (1989) and sample size tables for the Satterthwaite t in nQuery Advisor.

If non-normality is an issue, planning a large study or considering transformations as above may be helpful; another possibility is to plan to use a non-parametric procedure instead, such as the two-sample Mann-Whitney/Wilcoxon rank test. For a description of this test, see Rosner (2000), and for methods to determine sample size and power, see Hettmansperger (1984) and Noether (1987), or sample size tables in nQuery Advisor.

Note that the sample size methods for comparisons of two independent means discussed above do not apply to correlation/agreement/repeated measures (or pair-matched case-control) studies in which N subjects are recruited and each subject is measured by two different raters, or is studied under two different treatments in a crossover design. These designs cannot be analyzed using the methods described for independent means but must be analyzed using the paired t test or a repeated measures analysis of variance; see Rosner (2000) for information on the paired t test and Muller and Barton (1989) or sample size tables in nQuery Advisor for information about sample size and power for repeated measures tests.

28.6 Logistic Regression Models

In prior sections of this chapter, we discussed sample size problems for estimation or testing of a proportion in one or two groups. In this section, we consider study designs in which it is planned to evaluate several predictor variables for a binary outcome variable. Specifically, we consider studies in which we plan to fit a logistic regression model which is one of the most common tools in epidemiology. Readers needing an introduction to the logistic regression model and test procedures should consult Hosmer and Lemeshow (2000) (see also chapter ▸Regression Methods for Epidemiological Analysis of this handbook).

In our experience, there are two sample size questions, prospective and retrospective. The prospective question is: How many subjects do I need to observe to test the significance of a specific predictor variable or set of variables? The retrospective question is: Do I have enough data to fit this model? In this section, we consider methods for choosing a sample size first and then discuss the importance of having an adequate number of events per covariate.

With respect to planning sample size for logistic regression, we distinguish two situations: (1) only a single covariate is of interest, and (2) the addition of one covariate to a model already containing k covariates is of interest. In addition, we must distinguish whether the covariate of interest is dichotomous or continuous.

The basic sample size question is as follows: What sample size does one need to test the null hypothesis that a particular slope coefficient, say for covariate 1, is equal to zero versus the alternative that it is equal to some specified value?

28.6.1 Single Dichotomous Covariate

If the logistic regression model is to contain a single dichotomous covariate, then one may use conventional sample size formulas based on testing for the equality of two proportions. Hsieh et al. (1998) recommend using the following method to obtain sample sizes for logistic regression with a dichotomous covariate. (Although Whitemore (1981) provides a sample size formula for a logistic regression model containing a single dichotomous covariate, this formula, based on the sampling distribution of the log of the odds ratio, was derived under the assumption that the logistic probabilities are small and may be less accurate than the method we outline.)

Let the covariate X define two groups: group 1 contains those subjects for which $x = 0$ and the probability that the outcome of interest $y = 1$ for the subjects in this group is π_1, while group 2 contains those subjects for which $x = 1$ and the probability that $y = 1$ for these subjects is π_2.

Example 28.14.
Suppose that about 1% of the population is expected to have a particular adverse reaction to a certain drug used to treat a severe illness. It is thought that those with a specific preexisting condition (expected to be about 20% of the population) will be much more likely to have such a reaction; it will be important to detect an odds ratio of 2 for the likelihood of a reaction in this group using a 5% two-sided likelihood ratio test.

Table 28.15 Two-group χ^2 test of equal proportions (odds ratio = 1) (unequal n's)

Test significance level, α	0.050	0.050	0.050
1- or 2-sided test?	2	2	2
No condition proportion, π_1	0.010	0.010	0.010
Preexisting proportion, π_2	0.020	0.029	0.039
Odds ratio, $\psi = \pi_2(1 - \pi_1) / [\pi_1(1 - \pi_2)]$	2.000	3.000	4.000
Power (%)	90	90	90
n_1	7,620	2,468	1,345
n_2	1,905	617	337
Ratio n_2/n_1	0.250	0.250	0.250
$N = n_1 + n_2$	9,525	3,085	1,681

To compute the required sample size for Example 28.14 by hand would require using a modification of Eq. 28.27 for comparison of two proportions with unequal sample sizes, see references given in that section. Table 28.15 shows the table of results pasted from nQuery Advisor. (In this table, the symbol ψ is used to denote the odds ratio.) Defining group 1 as those without the preexisting condition and group 2 as those with, the ratio of the sample size in group 2 to the sample size in group 1 will be $20/80 = 0.25$. Using $\pi_1 = 0.01$ for group 1 (no preexisting condition), and $OR = 2$, we find $\pi_2 = 2\,(0.01)/(2\,(0.01) + 0.99) = 0.02$. Table 28.15 shows that to detect an odds ratio of 2 with 90% power for this example would require a sample size of 9,525. Consequently, the investigator may be interested in looking at the sample sizes required to detect odds ratios of 3 or 4 (3,085 and 1,681, respectively).

28.6.2 Single Continuous Covariate

If the single covariate we plan to include in the model is continuous, approximate formulas for this setting have been derived by Hsieh (1989) and implemented in sample size software packages such as nQuery Advisor. However, Hsieh et al. (1998) demonstrate that this approximate formula gives larger than required sample sizes and recommend using the following method to obtain sample sizes for logistic regression with a continuous covariate.

Let the response Y define two groups: group 1 contains cases in which $Y = 1$ with $N\pi_1$ cases expected, while group 2 contains cases in which $Y = 0$ with $N(1 - \pi_1)$ cases expected. The ratio of the expected sample size in group 2 to the expected sample size in group 1, r, is $(1 - \pi_1)/\pi_1$. The natural log of the odds ratio, the coefficient β of the covariate, x, is equal to the difference between the mean of the covariate in group 1 and the mean of the covariate in group 2 divided by the within-group standard deviation of x (denote this by δ). Therefore, a sample size formula or table for the two-group t test with unequal n's can be used to estimate sample size for logistic regression with one continuous covariate.

Example 28.15.
Patients with blocked or narrowed coronary arteries may undergo interventions designed to increase blood flow. Typically, about 30% of patients followed for a year will have renewed

blockage, called "restenosis," of the artery. A study is to be planned to use logistic regression to assess factors related to the likelihood of restenosis. One such factor is serum cholesterol level. Based on the results of a large screening trial, mean serum cholesterol in middle-aged males is about 210 mg/dL; one standard deviation above the mean (which corresponds to about the 85th percentile) is approximately 250 mg/dL. In the screening study, the odds ratio for the 6-year death rate for these two cholesterol levels was about 1.5. The study should be large enough to detect an effect of serum cholesterol on arterial restenosis of a size similar to that seen for death rate. We plan to conduct the test of the predictive effect of cholesterol level on the probability of restenosis using a 5% two-sided test and want to have 90% power to detect an odds ratio of 1.5 for values of cholesterol of 250 mg/dL versus 210 mg/dL. We set the effect size, $\delta = (\mu_1 - \mu_2)/\sigma = 0.405$, which is the value of the natural log of the odds ratio, 1.5. The ratio of sample sizes expected to be in the no-restenosis versus the restenosis groups, r, equals 0.7/0.3 = 2.333.

The required sample size could be computed using a version of Eq. 28.42 modified for unequal sample sizes, see references in the preceding section. In Table 28.16, we show the table of results pasted from the software nQuery Advisor.

To obtain a power of 90% to detect an odds ratio of 1.5 using the covariate cholesterol to predict restenosis at 1 year, we find that a total sample size of 310 is required.

28.6.3 Adjusting Sample Size for Inclusion of *k* Prior Covariates (Variance Inflation Factor)

It is rare in practice to have final inferences based on a univariate logistic regression model. However, the only sample size results currently available for the multivariable situation are based on very specific assumptions about the distributions of the covariates. We can, however, use a "variance inflation factor" to adjust the sample size results obtained for a single covariate for the situation in which k covariates have already been added to the model before the new covariate is considered.

The sample size, N_k required to test for the significance of a covariate after inclusion of k prior covariates in the model, is given by

$$N_k = N\left(\frac{1}{1-\rho^2}\right), \tag{28.44}$$

Table 28.16 Two-group t test of equal means (unequal n's)

Test significance level, α	0.050
1- or 2-sided test?	2
Effect size, $\delta = \lvert\mu_1 - \mu_2\rvert/\sigma$	0.405
Power (%)	90
n_1	93
n_2	217
Ratio n_2/n_1	2.333
$N = n_1 + n_2$	310

where the factor $1/\left(1-\rho^2\right)$ is called the "variance inflation factor"

$$VIF = \left(\frac{1}{1-\rho^2}\right) \tag{28.45}$$

and ρ^2 is the squared multiple correlation of the covariate of interest with the covariates already included in the model. This can be estimated using any multiple regression software.

For Example 28.14, the total sample size was computed as $N = 1{,}681$ for testing the significance of one dichotomous covariate. Now, assume that four patient demographic variables will be entered into the logistic regression model prior to testing the covariate indicating presence or absence of the preexisting condition (x_1 say) and that these demographic variables have a squared multiple correlation with x_1 of 0.2. Then a total sample size of at least 2,100 patients would be required:

$$N_4 = 1{,}681\left(\frac{1}{1-0.2}\right) = 2{,}101.$$

In Example 28.15, if two other covariates with a squared multiple correlation with cholesterol of 0.15 are to be entered into the logistic regression first, multiply the sample size obtained for a single covariate by the variance inflation factor, Eq. 28.44, $1/\left(1-\rho^2\right) = 1.18$, to increase the required sample size to 365.

28.6.4 Assessing the Adequacy of Data Already Collected

So far, we have discussed planning what sample size should be obtained to fit specific logistic regression models. A second consideration, and one relevant to any model being fit, is the issue of what is the maximum number of covariates it is reasonable to enter into the model and still obtain reliable estimates of the regression coefficients and avoid excessive shrinkage when the model is assessed for new cases. An ad hoc rule of thumb is to require that there be 10 "events" per variable included in the model. Here, the "event" of relevance is the least frequent of the outcomes. For example, suppose the study discussed in Example 28.15 was planned with 365 cases. Further, suppose that complete 1-year follow-up was only obtained for 351 cases of which 81 had restenosis at 1 year. There are 81 cases with restenosis and 270 without, so the least frequent "event" is restenosis. Based on these 81 cases, only 8 variables should be fit; this means that no more than 8 covariates (or covariates plus covariate interaction terms) should be entered into the model.

This rule of thumb was evaluated and found to be reasonable by Peduzzi et al. (1996) using only discrete covariates. However, as is the case with any overly simple solution to a complex problem, the rule of 10 should only be used as a guideline and a final determination must consider the context of the total problem. This includes the actual number of events, the total sample size, and most importantly, the mix of discrete, continuous interaction terms in the model. The "ten events per parameter" rule may work well for continuous covariates and discrete covariates with a balanced distribution. However, its applicability is less clear in settings where the distributions of discrete covariates are weighted heavily to one value.

28.7 Practical Issues in Sample Size Choice

In earlier sections, we outlined formulas for sample size computation for estimation and testing in simple designs for proportions and for means. We have shown only formulas to compute sample size from specifications of confidence interval width or desired power, but it is also possible to compute the confidence interval width or power which would be obtainable with a specified sample size. Sample size methods exist for many more complex designs and for other parameters. Software such as nQuery Advisor (Elashoff 2002) can be helpful.

For complex study designs or complex statistical methods, however, there may be no easily applied formulas or available software solutions. In such cases, sample size choices may be based on simplifications of the design or statistical methods (as we illustrated in the section on logistic regression), or in some cases a simulation study may be warranted.

For studies involving complex survey designs, sample size computations might be based on one of several approaches: (1) regarding the cluster itself as the study "subject" and using intraclass correlation values to estimate the appropriate variance to use in making computations, (2) multiplying sample sizes for a simpler design by a computed "design effect" (2 may be a sensible ad hoc choice), or (3) using simulation methods.

Although study sample sizes are usually chosen to assure desired precision or power for the primary outcome variable, investigators may also need to investigate whether that sample size choice will be adequate for evaluations of secondary outcomes, or for analyses of predefined subsets.

Sample size values obtained from formulas or software will generally need to be inflated to allow for expected dropout or loss to follow-up of study subjects or other sources of missing data (cf. chapter ▸Missing Data of this handbook). It is important to remember, however, that subjects who drop out may not be similar to those remaining in the study. This consideration may affect the parameter values which should be used for sample size computations, and even analyses using missing data techniques may not remove biases due to dropout.

Anotherissue of great concern to epidemiologists is that exposure or response may be misclassified (cf. chapter ▸Misclassification of this handbook). Such misclassification might have a dramatic impact on the actual power of a planned study unless sample sizes are computed based on modeling the expected type and extent of misclassification using simulation methods.

For brevity, our examples used only one set of parameter values to compute required sample sizes. In practice, investigators need to keep in mind that the estimated parameter values used in computations are only estimates and perhaps not very accurate ones. It is a good idea to compute necessary sample size for several different sets of parameter choices to evaluate sample size sensitivity to varying realistic possibilities for the true parameter values. Tables and plots can be helpful in these evaluations.

Finally, sample size justification statements in protocols, grant proposals, and manuscripts need to be complete. Details of the outcome variable, the study design, the planned analysis method, confidence level or power, one- or two-sided, and all the relevant distributional parameters (proportions, means, standard deviations) need to be included in the statement. For Example 28.13, a minimal sample size justification might read as follows: *A sample size of 19 in each group will have 90% power to detect a difference in means of 2.2 (the difference between an active drug mean change in hematocrit of 2.2% and a placebo mean change of 0.0) assuming that the common standard deviation is 2.0 and using a two-group t* test *with a 0.05 two-sided significance level. The planned enrollment will be 25 subjects per group (50 total) to allow for 20% dropout.* It is also desirable to provide information about sample size for other parameter choices and details about how these parameter values were selected, including references to previous studies which were consulted in selecting the values.

28.8 Conclusions

An important part of planning any research study is to assess what sample size is needed to assure that meaningful conclusions can be drawn about the primary outcome. To do this, the investigator must detail the study design, define the primary outcome variable, choose an analysis method, and specify desired or expected results of the study. Then formulas, tables, and sample size software of the sort outlined in this chapter can assist with computations. The most essential part of the process, though, is to make a thorough investigation of other information and research results concerning the outcome variable to support reasonable specification of hypothesized values for use in making computations. Beginning investigators often protest: “But this study has never been done before; how do I know what the results will be?” In most cases, however, much information about rates, means, and standard deviations can be gleaned from other contexts and used to infer what kinds of outcomes would be important to detect or likely to occur. Sample size computations are not just a pro forma requirement from funding agencies but provide the basis for deciding whether a planned study is likely to be worth the expense.

References

Chernick MR, Liu CY (2002) The saw-toothed behavior of power versus sample size and software solutions: single binomial proportion using exact methods. Am Stat 56:149–155

Dixon WJ, Massey FJ (1983) Introduction to statistical analysis, 4th edn. McGraw-Hill, New York

Elashoff JD (2002) nQuery Advisor® Release 5. Statistical Solutions, Ireland

Fleiss JL (1981) Statistical methods for rates and proportions, 2nd edn. Wiley, New York

Fleiss JL, Tytun A, Ury SH (1980) A simple approximation for calculating sample sizes for comparing independent proportions. Biometrics 36:343–346

Hettmansperger TP (1984) Statistical inference based on ranks. Wiley, New York

Hosmer DW, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley, New York

Hsieh FY (1989) Sample size tables for logistic regression. Stat Med 8:795–802

Hsieh FY, Bloch DA, Larsen MD (1998) A simple method of sample size calculation for linear and logistic regression. Stat Med 17:1623–1634

Korn EL (1986) Sample size tables for bounding small proportions. Biometrics 42:213–216

Kupper LL, Hafner KB (1989) How appropriate are popular sample size formulas? Am Stat 43:101–105

Lachin JM (1981) Introduction to sample size determination and power analysis for clinical trials. Control Clin Trials 2:93–113

Lachin JM (1992) Power and sample size evaluation for the McNemar test with application to matched case-control studies. Stat Med 11:1239–1251

Lemeshow S, Hosmer DW, Klar J, Lwanga SK (1990) Adequacy of sample size in health studies. Wiley, Chichester

Levy PS, Lemeshow S (1999) Sampling of populations: methods and applications, 3rd edn. Wiley, New York

Louis TA (1981) Confidence intervals for a binomial parameter after observing no successes. Am Stat 35:154

Machin D, Campbell MJ (1987) Statistical tables for design of clinical trials. Blackwell, Oxford

Machin D, Campbell M, Fayers P, Pinol A (1997) Sample size tables for clinical studies, 2nd edn. Blackwell Science, Malden/Carlton/London

Moser BK, Stevens GR, Watts, CL (1989) The two-sample *t* test versus Satterthwaite's approximate F test. Commun Stat Theory Methods 18:3963–3975

Muller KE, Barton CN (1989) Approximate power for repeated-measures ANOVA lacking sphericity. J Am Stat Assoc 84:549–555

Noether GE (1987) Sample size determination for some common nonparametric tests. J Am Stat Assoc 82:645–647

O'Brien RG, Muller KE (1983) Applied analysis of variance in behavioral science. Marcel Dekker, New York, pp 297–344

Peduzzi PN, Concato J, Kemper E, Holford TR, Feinstein A (1996) A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 99:1373–1379

Rosner B (2000) Fundamentals of biostatistics, 5th edn. Duxbury, Boston

Whitemore AS (1981) Sample size for logistic regression with small response probability. J Am Stat Assoc 76:27–32

Analysis of Continuous Covariates and Dose-Effect Analysis

29

Heiko Becher

Contents

H. Becher
Unit of Epidemiology and Biostatistics, Ruprecht Karls-University Heidelberg, Institute of Public Health, Heidelberg, Germany

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_16,

29.1 Introduction

When analyzing data from an epidemiological study, some features are rather specific for a particular study design. Those are dealt with among others in chapters ▶Descriptive Studies, ▶Cohort Studies, to ▶Modern Epidemiological Study Designs and ▶Survival Analysis of the handbook. Other features are generally relevant; see chapters ▶Rates, Risks, Measures of Association and Impact and ▶Confounding and Interaction. This chapter focusses on the analysis of continuous covariates where it will be discussed how such variables can be modeled to capture their potential association with an outcome of interest and to best describe the shape of such an association. We present classical methods based on categorization and subsequent contingency table analysis. The major part of the chapter, however, deals with the analysis of continuous covariates using regression models commonly used in epidemiology (see also chapter ▶Regression Methods for Epidemiological Analysis of this handbook). Each of the proposed techniques to model continuous covariates is illustrated by a real data example taken from a case-control study on laryngeal cancer and smoking as well as alcohol consumption that has been conducted in Germany during the 1990s. The chapter ends with practical recommendations and conclusions.

When planning the data analysis, we may distinguish between the following:

- Descriptive analysis
- Analytical statistical modeling

where the descriptive part may precede the analytical part. A thorough and careful descriptive analysis of the data can save enormous time. This involves univariate analysis of all variables of interest by graphical or other methods as described in textbooks on descriptive statistics (e.g., Bernstein and Bernstein 1998).

In practice, the analysis step is often started before the descriptive analysis has been completed. This is sometimes driven by impatient clinicians who would like to know immediately whether "the result is significant," that is, directly after data collection is finished and before plausibility checks of the data have been performed (see chapter ▶Data Management in Epidemiology of this handbook). However, the first p-value generated in the analysis step is rarely found in the publication. Instead, some peculiarities in the data are found such that one has to go back to the descriptive analysis. Then, we have an iterative process which should be made explicit. If used in an exploratory approach, an iterative process could provide a deeper insight and lead to general conclusions that may have not been envisioned at the outset.

In this chapter, we give an overview of methods to analyze continuous covariates in epidemiological studies. For a long time, it has been common to categorize them into $K \geq 2$ groups and to estimate appropriate parameters (e.g., odds ratio or relative risk) that describe the effect for a certain level of the variable relative to an arbitrarily defined baseline level or to perform tests for trend based on this categorization. This procedure was exclusively used in the past because classical methods to analyze epidemiological data, for example, the Mantel-Haenszel estimator, are mostly based on some form of contingency table analysis that requires a categorization of continuous variables. In the last decades, they have largely been replaced by regression models that allow, but do not require, a categorization. With increasing availability of powerful computers and standard software, these methods have become most common.

We will describe and compare different methods for the analysis of continuous covariates in classical epidemiological studies (such as case-control or cohort studies) and for the analysis of morbidity and mortality data. The aim is either (i) to analyze the relation between the outcome variable Y and the continuous covariate X or (ii) to appropriately adjust for X if one wishes to analyze the relation between Y and Z, where Z is another covariate which may be confounded by X (see also chapters ▸Basic Concepts and ▸Confounding and Interaction of this handbook). For (i), this is commonly done in epidemiology by estimating a relative risk or odds ratio function $\mathrm{R}(\cdot)$, where $(\cdot)$ denotes dependence on an arbitrary argument. For (ii), one has additionally the possibility to condition on X.

We will also consider the common situation that a proportion of individuals have exposure zero, and among those exposed, we have a continuous distribution of X. This is called a semicontinuous variable (Olsen and Schafer 2001), a spike at zero, (Robertson et al. 1994; Schisterman et al. 2006), mass at zero, or clump of zeros. Typical examples are occupational exposures, for example, asbestos exposure, or alcohol and tobacco consumption where a proportion of individuals may be completely unexposed, and the exposure of those who had been exposed follows a continuous (positive) distribution.

29.2 Classical Methods of Analyzing Continuous Covariates Based on Contingency Tables

Before the theory of generalized linear models was developed, contingency table analysis was the common tool in our field. The famous paper by Mantel and Haenszel (1959) is still well known, mainly because the ideas developed therein facilitated the path to modern techniques. In this section, we will therefore outline odds ratio estimation, tests for trend, and simple confounder adjustment approaches based on these classical methods. Although less often used today, these methods are still helpful for a general understanding of epidemiological data analysis. It should be emphasized, however, that the methods presented here have been augmented by modern regression methods which are now to be preferred.

29.2.1 General Aspects

We assume that a continuous covariate X has been collected and that Y is a binary variable that denotes the disease or the event of interest. The classical methods first require a categorization of X which can formally be described by the function

$$f(x) = \sum_k x_k I_{x \in I_k}(x),$$

where I is the indicator function being 1 if x belongs to I_k, which are disjunct exposure categories with $k = 1, \ldots, K$ covering the whole set of possible values of X, and 0 otherwise. The indicator functions are then multiplied with values x_k where a common method is to choose $x_k = k$, such that the kth largest group is assigned the value k, but it is also possible to assign x_k the mean exposure level in category k.

The practical questions are:

(i) How to choose the number of categories K
(ii) How to choose the intervals I_k

For neither question exists a unique or optimal answer. Since categorization is an important method also in regression modeling as described later, the final result obtained from the same dataset will most likely differ, if two statisticians use different categories when analyzing the same dataset.

Some general rules may help as a guideline for building categories:

- Choose the limits and the number of categories such that a comparison with earlier studies is possible.
- If a natural baseline exists (e.g., “0” for non-exposed) and a low exposure is a priori regarded as relevant (e.g., smoking), choose the natural baseline (in the example: the non-smokers) as a separate category.
- K should not be larger than 10, but for small datasets, fewer categories are to be preferred. Three to six categories are most often used in practice.
- Choose the limits and number of categories a priori, that is, do not base this choice on risk estimates derived from the same data.

A categorization by quartiles or quintiles of the distribution of the variable is often done. This is statistically correct, but then comparison with other studies is somewhat hampered since the limits follow the exact observed distribution of the covariate. It is perhaps better to report “Individuals whose diastolic blood pressure is above 90 mm Hg have a risk of … to get disease D” than to say “Individuals in the upper quartile of the diastolic blood pressure have a risk of … to get disease D.” Again referral to previous studies may help to derive meaningful cut-offs rather than quantile values.

29.2.2 Odds Ratio Estimation

Assume now a categorization has been performed yielding a 2xK table as displayed in Table 29.1. Crude odds ratio estimation for a given exposure level k is done by

$$\widehat{OR}_k = \frac{a_k c_1}{c_k a_1}, k = 2, \ldots, K,$$

Table 29.1 A 2xK table with, for example, the first exposure level chosen as reference category

		Cases	Controls	Total
Exposure level	x_1	a_1	c_1	m_1
	x_2	a_2	c_2	m_2
	…	…	…	…
	x_K	a_K	c_K	m_K
Total		n_1	n_0	N

Table 29.2 Alcohol consumption in males; laryngeal cancer case-control study; Germany

		Cases		Controls			
		a_k	(%)	c_k	(%)	$\widehat{OR}$	95% CI
Alcohol intake	<= 25	57	24.2	303	43.2	1.0	
(g ethanol/day)	25− <=50	51	21.6	169	24.1	1.60	(1.05–2.45)
	50− <=75	39	16.5	113	16.1	1.83	(1.16–2.91)
	75+	89	37.7	117	16.7	4.04	(2.72–6.00)
Total		236	100.0	702	100.0		

and asymptotic 95% confidence intervals are given as

$$\left(\exp(\ln(\widehat{OR}_k) - 1.96\sqrt{\hat{\text{var}}(\widehat{OR}_k)}), \quad \exp(\ln(\widehat{OR}_k) + 1.96\sqrt{\hat{\text{var}}(\widehat{OR}_k)})\right)$$

with

$$\hat{\text{var}}(\ln(\widehat{OR}_k)) = \frac{1}{a_k} + \frac{1}{c_1} + \frac{1}{a_1} + \frac{1}{c_k}.$$

(see Example 29.1). The odds ratios can be displayed graphically such that the exposure level within each group is plotted on the x-axis, and the odds ratio on the y-axis. This is a useful first presentation of the results.

Example 29.1.
As an example, we consider data from a laryngeal cancer case-control study where the variable X, daily alcohol consumption, is categorized into four group as depicted in Table 29.2.
In this study, average alcohol intake 10 years before interview was assessed by asking for the frequency and amount of the consumption of different types of alcoholic beverages (beer, wine, liqueur, spirits). Average alcohol content of these beverages was then used to calculate the daily ethanol intake. The second last column gives the unadjusted (i.e., not adjusted for confounders, such as smoking) odds ratio in comparison with the baseline category (see chapter ►Rates, Risks, Measures of Association and Impact of this handbook for an introduction of the Mantel-Haenszel estimate of the adjusted odds ratio), and the corresponding 95% confidence intervals as obtained from the equations above are added in the last column. We observe an increasing odds ratio with increasing dose.

The classical methods to analyze continuous covariates based on contingency tables have become less and less common in the recent literature. This is because:

(i) A categorization necessarily results in a loss of information.
(ii) The classical analysis is embedded in regression models as a special case.
(iii) The possibility to adjust for multiple confounders is limited.
(iv) There are only limited options for a dose-response analysis.

29.2.3 Test for Trend

Instead of presenting an effect measure for separate categories, one is often interested in the general question whether the effect increases (or decreases) with increasing dose. Very often, the presentation of the results goes with the presentation of a “test for trend.” Although this is sometimes more specifically referred to as “test for linear trend,” this is not sufficient, since there is no unique “test for (linear) trend,” and it often remains unclear what the authors of a paper actually have done. Moreover, such a statistical test may yield a significant result indicating a trend even if the single risk estimates are only increased in the lower dose range but approach baseline risk at higher doses (Maclure and Greenland 1992). Also, if there is a non-monotonic relation between exposure and disease, does this mean that there is no trend?

Based on Table 29.1, two common tests for trend can be derived: The first test to be introduced here is the Mantel-Haenszel χ^2-test for linear trend which is based on the following test statistics:

$$\chi^2 = \frac{N^2(N-1)\left[\sum_{k=1}^{K} x_k(a_k - e_k)\right]^2}{n_1 n_0\left[N\sum_{k=1}^{K} x_k^2 m_k - \left(\sum_{k=1}^{K} x_k m_k\right)^2\right]},$$

where $e_k = \mathrm{E}(a_k) = \frac{m_k n_1}{N}$ is the expected number of cases in category k.

Under the null hypothesis of no trend, this test statistic has an asymptotic χ^2-distribution with 1 degree of freedom (d.f.). As typical for χ^2-test statistics, it compares the observed number in each category with the one expected under the null hypothesis where the x_k serve as weights. To achieve an asymptotic χ^2-distribution, the resulting squared sum of weighted differences in the nominator has to be appropriately standardized by accounting for the weighted sample sizes for each exposure level $x_k m_k$, the numbers of cases n_1 and controls n_0, as well as the total number of subjects N.

The second test to be mentioned here is the Cochran-Armitage test for trend which is based on the test statistic,

$$T = \frac{\left[\sum_{k=1}^{K} x_k(a_k n_0 - c_k n_1)\right]}{\sqrt{\frac{n_0 n_1}{N}\left[\sum_{k=1}^{K} x_k^2 m_k(N - m_k) - 2\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} x_i x_j m_i m_j\right]}},$$

that is asymptotically standard normally distributed under the null hypothesis.

In practice, the weights x_k are often chosen as k. This is appropriate when the differences of the mean levels between two adjacent categories are similar, which is, for instance, fulfilled if the categories are equidistant. If this is not the case, the result of a trend test can highly differ whether x_k (e.g., defined as the midpoint of the exposure category) or simply k is used in the formula above. For the last, open-ended category, there is no unique best solution to assign a weight x_k. Given that the rank k is not used as weight, the best solution would be to investigate the distribution of the variable and to use as x_k the expected value given the observations are larger than the lower limit of this category. Since this may be difficult, a quick and pragmatic solution is to take the lower limit of this category multiplied by 1.5. If few observations are in this category, the value used for x_k is not crucial. In any case, however, an exact description of the method used is important (see Example 29.2).

Example 29.2.
We consider the data from the case-control study on laryngeal cancer as presented in Table 29.2. For the Mantel-Haenszel χ^2-test, we have $\chi^2 = 48.51$, and for the Cochran-Armitage test, we get $T = 6.97$ with $x_k = k$ which are both highly significant ($p < 0.001$). Thus, our observation of an increasing odds ratio with increasing dose is supported by both statistical tests.

The underlying principle is identical for contingency table analysis and for regression models, but for the latter, there exists a larger variety of methods to perform a trend test (Agresti 2002).

29.2.4 Concluding Remarks on Classical Methods

In summary, methods for analyzing contingency tables, which have mainly been developed before 1980, have served as analytical techniques for analyzing epidemiological studies for a long time. For a more detailed treatment of such methods, we refer to classical textbooks like Breslow and Day (1980). It has become apparent, however, that these methods have their limits, as best seen in the analysis of continuous covariates. The major drawbacks are:

- The need to categorize continuous covariates, leading to a loss of information
- Limited possibilities to derive a dose-response curve
- Limited ability for confounder adjustment

In addition, the interpretation of the results may vary with the choice of the cut-points of the categories.

Nevertheless, these methods have one general advantage which should not be underestimated: they are simple, and some can easily be calculated with a small pocket calculator. As a first step, following a descriptive analysis of the data, they are still useful although these results will rarely find their way into high ranking journals today.

In the following, the most relevant regression models in epidemiological research are briefly summarized (see also chapter ▶Regression Methods for Epidemiological Analysis of this handbook) before several methods to analyze continuous covariates are presented.

29.3 Regression Models and Risk Functions

It has already been mentioned several times that regression modeling is the method of choice for the analysis of epidemiological data. The advantages of these are particularly apparent when continuous covariates have to be analyzed since they allow the analysis of the data without categorization. In this section, we introduce relevant regression models and a notation which is useful for the subsequent presentation of the methods to analyze continuous covariates.

Let Y be the outcome variable, and let X and Z be two covariates (X continuous, Z unspecified). In this chapter, we consider (a) the logistic regression model, where the outcome variable Y is dichotomous, $Y = 1$ (diseased), $Y = 0$ (not diseased); (b) the Poisson regression model, where $\mu = D/PY$ is the outcome variable with D the observed number of events, for example, deaths, and PY (person-years) the observation time for all individuals; and (c) the Cox regression model, where $\lambda(t)$ is the hazard function and one observes the individual survival time t and a censoring indicator (for a general discussion of regression models, see chapter ▶Regression Methods for Epidemiological Analysis of this handbook; for models of the latter type, we also refer to chapter ▶Survival Analysis). Although Cox regression is conceptually different from linear, logistic, or Poisson regression, the handling of continuous covariates follows the same principles. Therefore, these regression models are considered jointly here.

The commonly used (log-linear) form of the logistic, Poisson, or Cox model reads as follows for the logistic regression model:

$$P(Y = 1|x, z) = \frac{\exp(\beta_0 + \beta_1 x + \beta_2 z)}{1 + \exp(\beta_0 + \beta_1 x + \beta_2 z)}$$

the Poisson regression model:

$$\mu = \exp\left(\beta_0 + \beta_1 x + \beta_2 z\right),$$

and for the proportional hazards (Cox) regression model:

$$\lambda(t) = \lambda_0(t)\exp(\beta_1 x + \beta_2 z),$$

where β_1 and β_2 are the regression coefficients for X and Z, respectively, and β_0 is an intercept parameter. This standard form is not very flexible since a specific functional relation between the covariates and the outcome is fixed (see Sect. 29.4 for details). A more general form of these models allows transformations of X and Z with functions f and g, and the exponential function exp is replaced by R. The general form then becomes for the logistic regression model:

$$P(Y=1|x,z) = \frac{\exp(\beta_0)R_X(f(x),\beta_1)R_Z(g(z),\beta_2)}{1+\exp(\beta_0)R_X(f(x),\beta_1)R_Z(g(z),\beta_2)},$$

Poisson regression model:

$$\mu = \exp(\beta_0)R_X(f(x),\beta_1)R_Z(g(z),\beta_2),$$

and for the Cox regression model:

$$\lambda(t) = \lambda_0(t)R_X(f(x),\beta_1)R_Z(g(z),\beta_2),$$

where R_X and R_Z are risk functions.

For illustration, let us consider the logistic model first. In the classical form as given above, a dose-effect function for the variable X, given as $OR(X = x$ vs. $X = x_0)$, is given by

$$OR(X = x \text{ vs. } X = x_0) = \frac{P(Y=1|x,z)P(Y=0|x_0,z)}{P(Y=1|x_0,z)P(Y=0|x,z)} = \exp(\beta(x-x_0)),$$

thus automatically yielding a specific (exponential) form of the dose-effect curve. This is not sufficient for most analyses, and the more general form is given as

$$OR(X = x \text{ vs. } X = x_0) = \frac{P(Y=1|x,z)P(Y=0|x_0,z)}{P(Y=1|x_0,z)P(Y=0|x,z)} = R(\beta(f(x)-f(x_0))).$$

In the history of the development, different choices for R have been considered first (see Sect. 29.4.4). Transformations of the covariate X before entering it in the regression model have also been common for quite a while, but only after the development of fractional polynomials this has been investigated in a more systematic way (Royston and Altman 1994; Royston and Sauerbrei 2008; see section “Fractional Polynomials” in Sect. 29.4.3). The correct functional form of the relative risk function is commonly unknown, and unless biological knowledge can be added to support a specific shape, one has to base the decision on statistical grounds. For assessing the dose-effect curve associated with X, different approaches are used in the literature which are dealt with in the next section.

It is possible in some cases to analytically derive the appropriate model if the distribution of the covariates is known. If, for instance, X is normally distributed in cases and controls with equal variance and (possibly) different mean, then the classical form $R(\cdot) = \exp(\cdot)$ and $f = \mathrm{id}$ (where id denotes identity, i.e., $f(x) = x$) is the correct one. In practice, however, it is more complex and it is necessary to find a statistical procedure for building the model and for discriminating between models (Becher 1993; Becher et al. 2012).

It is noteworthy at this point that $R(\cdot) = \exp(\cdot)$ describes the odds ratio (or relative risk or rate ratio) for the value of $X = x$ relative to a reference level $X = x_0$. For example, if X denotes the number of cigarettes smoked, then $R(10)$ describes the odds ratio for smoking 10 cigarettes compared to smoking zero, but likewise, the odds ratio for smoking 20 cigarettes compared to 10. This is the reason why the odds ratio function often does not fit to the estimates obtained by categorization of X where the odds ratio for each exposure category is estimated in comparison to the same baseline category, a fact that is nicely described in Breslow and Day (1980, pp. 220f). They estimated the risk for esophageal cancer for different amounts of tobacco consumption. Compared to the baseline (non-smoker), even moderate smokers have a considerable risk (odds ratio about 4.3). The increase in risk with increasing dose, defined as four different amounts of tobacco consumption, is clearly not log-linear (see Fig. 6.4 in Breslow and Day (1980), p. 221). When estimating the risk with the classical logistic model, the odds ratio function is, however, forced to be log-linear with $\log(OR(X = x \text{ vs. } X = x_0)) = \beta(x - x_0)$ which does not fit the data well.

The risk function R may be embedded in the Poisson regression in a similar way as in logistic regression. However, there is one principal difference when applying this regression model. Here, a rate is modeled which results from the observed number of events in a category divided by the corresponding observed person-time (see chapter ▶Cohort Studies of this handbook). Therefore, a categorization of a continuous covariate is necessary for the allocation of person-time (see Breslow and Day 1987, pp. 85–86). For example, if the effect of average daily alcohol consumption X (in g ethanol per day) on a particular disease is analyzed from a cohort study with Poisson regression, then X has to be categorized first, for example, in categories 0, >0–10, >10–20, >20–40, >40–80, and 80+ g/day. Then, the observed number of cases and corresponding person-years are calculated for each exposure category, and finally the Poisson regression can be performed using one of the methods as described later. The dose-effect analysis is then based on an average exposure level within an exposure category for which the mean of all individuals falling into the respective category can be used.

29.4 Methods to Analyze Continuous Covariates in Regression Models

In this section, several approaches will be described to include continuous covariates in regression models where we will first briefly state the underlying principle of the respective approach before this will be further acknowledged and illustrated by applying it to the case-control study on laryngeal cancer.

29.4.1 Categorization of the Covariate into K Categories

Underlying principle: A continuous variable X is transformed into K binary variables. In the regression model, $K-1$ regression coefficients are estimated.

Here we use

$$R(\cdot) = \exp(\cdot);\ f(x) = (I_{x \in I_1}(x), \ldots, I_{x \in I_K}(x)),$$

where I is the indicator function and I_K are the exposure categories $1, \ldots, K$, as defined in Sect. 29.2.1.

This method is still commonly used in epidemiology. It corresponds to classical methods for the analysis of grouped data. In contrast to categorization of the continuous covariate and building contingency tables as described in Sect. 29.2, the covariate is split into K binary variables which then enter the regression model. These are also called "dummy variables." The category that defines the baseline, usually the non-exposed or the low exposed, is left out of the model. The criteria for choosing the number of categories and their limits are the same as for the classical methods described in Sect. 29.2.

This approach has the advantage that it leads to results that can be easily interpreted. It is therefore popular among non-statisticians, and it is "model-free" in the sense that it does not assume a specific shape of the dose-effect relationship. It is a useful step within the overall analysis procedure; however, it also has some serious drawbacks. Among these are an arbitrary choice of the baseline category, cut-points, and number of categories. The full information from the data is not used and the risk function is by definition a step function:

$$\mathrm{R}(x) = \exp\left(\sum_{k=1}^{K} \beta_{1k} I_{x \in I_k}(x)\right).$$

If X is a confounder for Z and if K is small (e.g., $K = 2$), a high degree of residual confounding is likely to occur. If K is large, residual confounding is less likely, but then a high number of parameters has to be estimated which may not be feasible if the sample size is small. Often, a categorization (dichotomization) of X is made such that the p-value is minimal. This procedure, though often found in literature, yields false results if this p-value is used without adjustment. Uncritical use of the minimal p-value approach would result in a marked increase in the false-positive error rate, and many factors that have no risk potential would be labeled erroneously as risk factors, increasing the existing problems in epidemiological research of simultaneous assessment of associations between many potential risk factors and disease outcome (Altman et al. 1994; Schulgen et al. 1994).

A "test for linear trend" which parallels the classical test introduced in Sect. 29.2.3 is given by assigning the values $1, 2, 3, \ldots, K$ to the K categories of X and entering X with these values into the regression model. The advantage

Table 29.3 Alcohol consumption in males; laryngeal cancer case-control study; Germany. Analysis with logistic regression, Method of Sect. 29.4.1

Parameter	Estimate	Standard error	Wald χ^2	p-value	$\widehat{OR} = \exp(\hat{\beta}_i)$	95% CI
α	-1.69	0.15	134.73	<0.0001		
β_2	0.47	0.22	4.70	0.0301	1.60	(1.046 − 2.449)
β_3	0.67	0.23	8.37	0.0038	1.96	(1.234 − 3.101)
β_4	1.41	0.20	48.89	<0.0001	4.12	(2.768 − 6.119)

over the methods described in Sect. 29.2.3 is the possibility to simultaneously adjust for other covariates which is only possible to a limited degree using the classical methods. The disadvantage, namely, the restriction to a specific functional form of the dose-response function, however, is the same as before, and therefore it is not recommended for general use.

This approach also works for variables with a spike at zero. In this case, the baseline category is typically given by the non-exposed ($X = 0$). Problems may arise if this category is sparse. In that case, as in the example below, a category "no or low exposure" may be more appropriate because otherwise the standard errors of regression coefficients become large.

In the following, we will illustrate the above approach by using the data from the case-control study presented in Table 29.2 where alcohol consumption is now coded as a dummy variable and modeled as potential risk factor for laryngeal cancer using a logistic model (see Example 29.3).

Example 29.3.
Using the data from Table 29.2, we can fit a logistic model with alcohol consumption coded as dummy variable with the lowest category $i = 1$ (<=25 g/day) as the baseline category. The model then reads as logit$(P(Y = 1|\text{alcohol level } i) = \alpha + \beta_i$, $i = 2, 3, 4$, and the odds ratio estimates OR(alcohol level i vs. alcohol level 1) are obtained by $\exp(\hat{\beta}_i)$. Using the above model with the SAS-statement

```
Proc logistic descending; model y = x₂ x₃ x₄;
```

where X_2, X_3, X_4 are binary variables which take the value 1 if the observation falls in the i-th category and 0 otherwise, we get the results given in Table 29.3 (adapted from the SAS output).

The result corresponds to the result from the classical method as given in Table 29.2.

29.4.2 Leaving the Covariate Untransformed

Underlying principle: Values of X are included into the regression model as recorded. One regression coefficient is estimated.

Thus, we get

$$f(x) = x; R(\cdot) = \exp(\cdot), \text{yielding } R = \exp(\beta x).$$

This is the standard approach which uses the full information of the measured variable. It can be applied also for variables with a spike at zero. However, it has the serious drawback that it assumes the exponential function as the dose-effect curve. This method is valid only in special situations, for example, if X has a normal distribution in cases and controls with equal variance. In many papers, one finds statements like "odds ratio estimates are adjusted for age" which often simply means that this method has been employed to adjust for X (age). The method can easily yield false results. In Example 29.8, we present data from a case-control study in which this method is not appropriate. The Method described in this section has a strong underlying model assumption, namely, that the risk function is log-linear. If this assumption is violated and the method is nevertheless applied, it has the following consequences: (i) the resulting dose-response curve does not represent the true relationship between the factor and the outcome and (ii) the regression estimates of other variables in the model are not appropriately adjusted for the effect of X.

Testing the null hypothesis of no effect of the exposure, that is, H_0: $\beta = 0$, the corresponding p-value for $\hat{\beta}$ as obtained from the model fit may be regarded as the result of a "test for linear trend." In that context, "linear" means that the log odds ratio increases linearly with dose. (see Example 29.4).

Example 29.4.
We illustrate Method 4.2 with the individual data on alcohol consumption, X, from the case-control study on laryngeal cancer. The model logit $P(Y = 1|x) = \beta_0 + \beta_1 x$ with the SAS-statement

```
Proc logistic descending; model y=x;
```

yields the results given in Table 29.4 (adapted from the SAS output).
Under the model assumption that the log odds ratio linearly depend on the alcohol dose, we get the dose-effect curve $OR(X = x* \text{ vs. } X = x_0) = \exp(0.00878(x * -x_0))$. A further evaluation whether this assumption is justified is not performed.

Table 29.4 Alcohol consumption in males; laryngeal cancer case-control study; Germany. Analysis with logistic regression, Method of Sect. 29.4.2

Parameter	Estimate	Standard error	Wald χ^2	p-value
α	−1.59	0.1073	220.93	<0.0001
β	0.0088	0.00125	49.22	<0.0001

29.4.3 Monotonic Transformations of the Covariate and Fractional Polynomials

Underlying principle: X is transformed before entering the regression model. Monotonic and non-monotonic dose-effect functions can be obtained.

Here, X undergoes a monotonic transformation with common functions, such as the logarithm or square root before it enters the regression model. This has been a common "ad hoc" method. Such monotonic transformations can be easily realized within the various software routines for calculating regression models.

Figure 29.1 shows the shape of the dose-effect curve for three common transformations

$$f_1(x) = \sqrt{x}, \quad f_2(x) = x^2, \quad f_3(x) = \log(x+1),$$

for which the regression coefficient is chosen such that $R(1) = \exp(1)$. It is seen that f_1 generates an almost linear dose-effect relationship within the given dose range, f_2 generates a strictly increasing (quadratic) curve, and f_3 a concave dose-effect curve (see also Sect. 29.4.4 for comparison).

For f_3, a specific feature must be noted: if the covariate X can be zero or negative, then the simple log transformation is not possible, and instead of $\log(x)$, $\log(x+k)$ has to be used (see also Sect. 29.4.5 below).

Different approaches have been proposed to discriminate between models as in Fig. 29.1 which are called "non-nested models" (e.g., Royston and Thompson 1995; Mizon and Richard 1986). These approaches are beyond the scope of this chapter. It is also possible to use a higher-dimensional transformation, for example, $f(x) = (x, x^2)$. Then, not only X is included into the model but also X^2. A formal test on whether the second component significantly improves the goodness of fit is readily available as the difference of deviances which is asymptotically

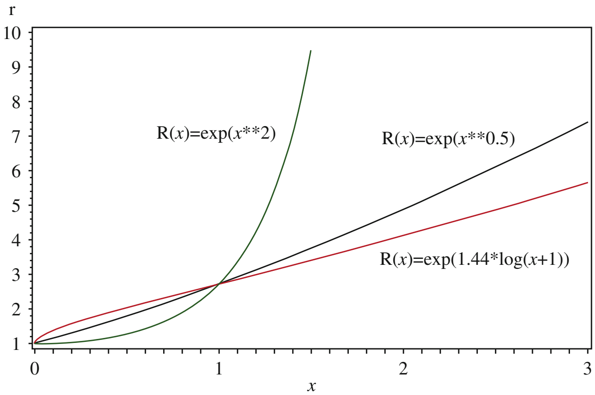


Fig. 29.1 Possible shapes of dose-effect curves as obtained from the transformations $f_1(x) = \sqrt{x}$, $f_2(x) = x$, $f_3(x) = \log(x+1)$

χ^2-distributed. A non-monotonic dose-effect function may result from this approach, if, for example, the regression coefficients have different signs.

An open problem so far is how the covariate should be transformed to achieve the best monotonic or non-monotonic dose-effect function in the sense of an optimal model fit. Royston and Altman (1994) have developed a stringent procedure which they called "fractional polynomials." A monograph by Royston and Sauerbrei (2008) gives a full account of the method which is further outlined below.

Fractional Polynomials The idea of fractional polynomials (FP) is to allow the variable to enter the model after it has been transformed where the transformation used is selected from a predefined set of eight different functions. This set is defined as $H_1(x) = x^p$ with $p \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ with x^0 being defined as $\log(x)$. In an FP of degree 1, the variable is entered successively using these eight transformations. In an FP of degree 2, the variable enters the model a second time with $H_2(x) = x^q$ with $q \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ and with $H_2(x) = x^p \log(x)$ in case that $q = p$. Royston and Altman (1994) showed that the fractional polynomials of degree 2 cover a very rich family of dose-effect relationships. Thus, a huge variety of dose-effect relationships can be fitted where often different FPs yield very similar fits with very similar dose-effect relationships (Royston and Altman 1994).

Embedding FPs in the above framework, we have $R(\cdot) = \exp(\cdot)$, and we use as linear predictor

$$\beta_0 + \sum_{j=1}^{J} \beta_j H_j(x) \quad ,$$

where, for example, for a first-order FP, J is equal to 1 and $H_1(x) = x^p$ as defined above. For a second-order FP (FP2), we have $J = 2$, and the models are of the form $\beta_0 + \beta_1 x^p + \beta_2 x^q$. FPs of higher order rarely occur in practice. In practice, $J = 2$ is most often used. If it can be assumed on biological grounds that the relation with the disease is monotonic, the FP method can be restricted to FPs of degree 1 ($J = 1$).

As already mentioned above, FPs were developed to provide a procedure to systematically search for the optimal transformation of the covariate. For this purpose, a closed test procedure, called FSP (function selection procedure), has been suggested which is described in detail in Royston and Sauerbrei (2008). This procedure consists of three steps which are described in the following:

Step 1: Test the best FP2 model for the covariate X at a significance level α against the null model (without the covariate X) using 4 d.f. If the test is not significant, the procedure has to be stopped concluding that the effect of X is "not significant" at level α. Otherwise, continue with Step 2.

Step 2: Test the best FP2 for X against a straight line (i.e., X remains untransformed) at significance level α using 3 d.f. If the test is not significant, the procedure has to be stopped concluding that the final model is a straight line. Otherwise, continue with Step 3.

Step 3: Test the best FP2 for X against the best FP1 at significance level α using 2 d.f. If the test is not significant, it can be concluded that the final model is FP1; otherwise, the final model is FP2 and the procedure ends.

The test used in Step 1 checks for an overall association of the outcome with the covariate X. The test at Step 2 examines the evidence for non-linearity in X. The test at Step 3 is used to select between a simpler or more complex model. All tests are based on the difference of deviances using the asymptotic χ^2-distribution.

In many practical situations (excluding the common spike at zero situation which is considered separately), the continuous exposure variable X is positive (e.g., blood pressure, vitamin A intake, body mass index). If X is an exposure variable which can take values smaller or equal zero (e.g., weight gain from age 20 to 40), they recommend to add a constant c to X in order to make all transformations possible. As an example for this case, consider as X the change in body mass index from age 20 to 30. This variable has a continuous distribution, and the observation range covers positive and negative values. In order to allow application of the FP procedure, a transformation $X \to X + c$ is required, where c has to be chosen being larger than the smallest observation of X, that is, $c > x_{\text{min}}$.

In general, please note that an FP of degree 1 yields monotonic dose-effect functions, whereas an FP of degree 2 allows for non-monotonic dose-response functions (see Example 29.5).

Example 29.5.

We illustrate the FP procedure with data of the case-control study on laryngeal cancer. We are interested in the effect of smoking dose among the smokers where the covariate smoking is considered as cumulative dose in pack-years. A SAS macro is available for this analysis. The analysis with a second degree FP gives the results shown in Table 29.5.

The best FP2 (where X is transformed to the power of 0.5 and 1) is compared with the null model at level $\alpha = 0.01$ using 4 d.f. The model identified at that stage is therefore logit $P(Y = 1|X = x) = \beta_0 + \beta_1\sqrt{x} + \beta_2 x$. We observe a difference of deviances of 257.46 which is highly significant. In the next step, we compare the best FP2 with the linear model at $\alpha = 0.01$ with 3 d.f. We get 35.69, which is also highly significant. Finally,

Table 29.5 Laryngeal cancer case-control study. Application of the standard FSP procedure to smoking dose (among smokers), $N = 766$

	Deviance	Deviance difference	d.f.	p-value	Power(s)
FP2	800.77		1	1.00	(0.5, 1)
Null		257.46	4	<0.001	–
Linear		35.69	3	<0.001	(1,–)
FP1		2.21	2	0.33	(0,–)

we compare the best FP2 with the best FP1 at $\alpha = 0.01$ using 2 d.f. The best FP1 is a polynomial of degree 0 which thus has the form logit $P(Y = 1|X = x) = \beta_0 + \beta_1 \log(x)$ (log transformation of the covariate). The result is 2.21 and non-significant. Therefore, the final result is an FP1 with the power 0. This is, in the notation of the FPs, the transformation $\log(\cdot)$. The estimated regression coefficient has the value 1.09, and the estimated dose-effect curve follows:

$$OR(X = x^* \text{vs. } X = x_0) = \exp(1.09 \times (\log(x^*) - \log(x_0))).$$

29.4.4 Additive (Linear) Risk Functions

Underlying principle: It is assumed that the risk function increases linearly with dose (as opposed to Sect. 29.4.2, where a log-linear risk function is assumed). One parameter is estimated.

In this case, we thus assume that

$$f(x) = x, R(x) = 1 + \beta x.$$

This risk function, though often described in the literature and also available in some software packages, is not very often used in practice. The method dates back to Breslow and Storer (1985). To get some guidance on which risk function should be used in a given situation, a goodness-of-fit statistic to compare the linear and exponential relative risk function, that is

$$R_1(x) = \exp(\beta x) \text{ vs. } R_2(x) = 1 + \beta x,$$

would be desirable. However, a simple test for the hypothesis, H_0: "Both models fit the data equally well" vs. H_1: "One model fits the data better than the other" is not available because the models are not nested and the difference of the deviances has an unknown distribution under the null hypothesis. The restriction $f(x) = x$ can easily be relaxed, that is, by a transformation of X with a function as given in Sect. 29.4.3, but this has also rarely been done in practice.

In this model, the risk when comparing two covariate values x_1 and x_2, $0 \leq x_1 < x_2$, does not only depend on $x_2 - x_1$ as it is the case for the log-linear risk function (see Sect. 29.4.2) since it holds

$$R(X = x_2 \text{ vs. } X = x_1) = \frac{1 + x_2\beta}{1 + x_1\beta} < 1 + (x_2 - x_1)\beta.$$

For example, let us assume $\beta = 1$. Then,

$$R(X = 1 \text{ vs. } X = 0) = \frac{1 + 1}{1 + 0} = 2,$$

whereas

$$R(X = 2 \text{ vs. } X = 1) = \frac{1+2}{1+1} = 1.5,$$

although the difference of the arguments is the same in both cases.

Note 1 A linear dose-effect curve is also obtained under $R(\cdot) = \exp(\cdot)$ by using $f(x) = \log(\beta x + 1)$ and choosing β such that $\gamma = 1$. It follows that $R(x) = \exp(\log(\beta x + 1)) = (1 + \beta x)$. This algebraic equivalence is used in the Example 29.6 below to fit an additive (linear) risk function with standard software, which usually has $R(\cdot) = \exp(\cdot)$ as default.

Example 29.6.
In this example, we illustrate the fit of an additive (linear) risk function of the form $R(x) = 1 + \beta x$. In order to use standard software, such as PROC LOGISTIC in SAS, the property shown in the note above is taken advantage of. We consider again the case-control study on laryngeal cancer and alcohol consumption as risk factor. We use the model $\text{logit}(X = x|Y = 1) = \alpha + \gamma \log(\beta x + 1)$ with $\beta = 0.02975$. We get a difference of deviances of 54.24 ($p < 0.001$) with 1 d.f. and a parameter estimate $\hat{\gamma} = 1$. Note again that β was chosen such that the regression parameter estimate $\hat{\gamma}$ becomes 1. As a linear risk function, we then get

$$OR(X = x^* \text{ vs. } X = x_0) = \exp(1 \times \log(0.02975x + 1)) = 1 + 0.02975x.$$

The inference for β is done by comparing the deviances of the models with and without the transformed variable X. The difference of deviances is asymptotically χ^2-distributed with 1 d.f. under the null hypothesis $\beta = 0$.

29.4.5 Spike at Zero

Underlying principle: X has a spike at zero which requires special consideration. X is transformed before it is entered the regression model according to a specified procedure. For this procedure a binary exposure indicator (exposed/nonexposed) is introduced. Monotonic and non-monotonic dose-effect functions can be obtained.

This method was motivated by the fact that there are covariates whose distribution has a discrete and a continuous component. The most prominent example is smoking where a certain proportion of the population are non-smokers referred to as $X = 0$, that is, $P(X = 0) > 0$. Among smokers, the distribution of smoking dose is continuous. In this situation, there are both statistical problems and difficulties to interpret the results. From the statistical point of view, the presence of zero exposure precludes some of the transformation, such as log or the inverse function. A suggested ad hoc procedure is to add a small constant to each observation, but this lacks theoretical justification. With regard to interpretation, the relative risk for smoking k cigarettes daily compared to 0 may not be the same as smoking $2k$ cigarettes compared to k.

Jedrychowski et al. (1992) took account of this problem by entering a binary variable smoker/non-smoker into the model in addition to the (transformed) dose variable (see examples). However, this was an ad hoc approach without giving a formal rationale for the procedure. The method was more formally described in Robertson et al. (1994). The first parameter then represents the effect of exposure, and the second parameter represents the effect of the categories of exposure among those who are exposed. But note that Robertson did not suggest a formal model selection procedure.

Recently, the spike at zero situation has been considered using an extended FP approach (Royston et al. 2010; Becher et al. 2012). The second paper by Becher et al. (2012) modifies the method by Royston et al. (2010) which leads to slightly better properties of this method. We therefore will only describe the modified version from Becher et al. (2012).

The function selection procedure (FSP) has four steps, which are given below:

1. Generate a binary indicator Z indicating whether the exposure X is present or not.
2. Fit the most complex model (binary indicator Z + 2nd degree FP) and compare the deviance of this model with the deviance of the null model using the χ^2-distribution with five degrees of freedom. If the corresponding test is not significant, stop the procedure concluding that the effect of the exposure is "not significant" at level α. The transformations are applied to the positive values of X only.
3. If significant, follow the same FP function selection procedure *with* Z included (first stage) as in the simple case without a spike.
4. Test both Z and the remaining FP (respective the linear component) for removal.

A more detailed description of the method is given in the following. The FSP-spike procedure for selecting a model has two stages. In the first stage (Steps 2 and 3), the most complex model comprising Z and $\text{FP2}+(X;p,q)$ is compared with the null model on 5 d.f. (4 d.f. from the FP2 model plus 1 d.f. from the binary Z term). If the test is significant, the steps of the FSP for selecting an FP function are followed, but with Z always included in the model. The result after this stage is either that an effect of the exposure cannot be shown (null hypothesis not rejected) or a model with a binary indicator plus an FP of degree 1 or 2.

In the second stage (Step 4), Z and the remaining FP or linear component are each tested for removal from the model. If both parts are significant, the final model includes both; if one or both parts are non-significant, the one with the smaller deviance difference is removed. In the latter case, the final model comprises either the binary dummy variable or the selected FP function. If only an FP function is selected, the spike at zero plays no further role and is removed.

Since the selection of an FP function may be affected by the presence of the binary dummy variable, the resulting model may differ from that of a standard FP analysis. An example for the approach is given in Example 29.7; further examples are given in Becher et al. (2012).

Example 29.7.
In this example, we illustrate the FSP-spike procedure with the extended FP approach. We are interested in the general effect of amount of smoking in comparison to non-smokers

Table 29.6 Laryngeal cancer case-control study. Application of the FP-spike procedure to smoking dose (smokers and non-smokers included), $N = 938$

	Deviance	Deviance difference	d.f.	p-value	Power(s)
First stage					
FP2 + spike	846.09		1	1.00	(0.5, 1)
Null		212.13	5	<0.001	–
Linear + spike		35.60	3	<0.001	(1,–)
FP1 + spike		4.95	2	0.08	(0,–)
Second stage					
FP1 + spike	851.04		3		
FP1 (dropping spike)		0.62	1	0.43	
Spike (dropping linear)		132.06	4	<0.001	

in the case-control study on laryngeal cancer with the exposure variable cumulative dose in pack-years. The general model has the form logit $P(Y = 1|X = x) = \beta_0 + \beta_1 z + \beta_2 x^p + \beta_3 x^q$, where $z = I_{[X=0]}(x)$ represents the binary indicator, p and q represent the powers in the FP model, and β_1 is the coefficient for the binary indicator variable. The analysis with a second-degree FP-spike gives the results shown in Table 29.6.

In the first stage, we identify the best FP2-spike model among all FPs. This model has powers 0.5 and 1 and includes the spike. The model identified at that stage is therefore logit $P(Y = 1|X = x) = \beta_0 + \beta_1 z + \beta_2 \sqrt{x} + \beta_3 x$. It has a deviance of 846.09. We compare this model with the null model at the level $\alpha = 0.01$ using 5 d.f. We observe a difference of deviances of 212.13 which is highly significant ($p < 0.001$) and conclude that the exposure has a significant effect on the outcome variable.

In the next step we compare the best FP2-spike with the linear model with spike at $\alpha = 0.01$ with 3 d.f. We get a difference of deviances between the best model and the linear model of 35.60 which is also highly significant. We conclude that the model logit $P(Y = 1|X = x) = \beta_0 + \beta_1 z + \beta_3 x$ is not sufficient to describe the data. Finally, we compare the best FP2-spike with the best FP1-spike at $\alpha = 0.01$ using 2 d.f. The best FP1-spike is a polynomial of degree 0 (log transformation) and has the form logit $P(Y = 1|X = x) = \beta_0 + \beta_1 z + \beta_2 \log(x)$. The difference of deviances between this model and the above best FP2-model is 4.95. This difference is not significant ($p = 0.08$), and we select the model logit $P(Y = 1|X = x) = \beta_0 + \beta_1 z + \beta_2 \log(x)$ in the first stage. This model has a deviance of 851.04.

In the second stage, we check whether either the spike or the selected FP1 can be omitted. Removal of the spike yields a slightly worse fit with a difference of deviances of 0.62. The spike variable is not significant ($p = 0.43$). The selected FP1, however, is highly significant with a difference of deviances of 132.06. This results from comparison of the model logit $P(Y = 1|X = x) = \beta_0 + \beta_1 z + \beta_2 \log(x)$ with the model logit $P(Y = 1|X = x) = \beta_0 + \beta_1 z$. Therefore, the final result is a FP1 of the form logit $P(Y = 1|X = x) = \beta_0 + \beta_2 \log(x)$. The resulting regression coefficient taken from this model is $\hat{\beta}_2 = 0.926$ and yields the estimated dose-effect curve:

$$OR(X = x^* \text{vs. } X = x_0 | x_0 > 0) = \exp(0.926 \times (\log(x^*) - \log(x_0)))$$

and

$$OR(X = x^* \text{vs. } X = 0) = \exp(0.926 \times \log(x^*)) = (x^*)^{0.926}.$$

We can conclude that the estimated odds ratio increases almost linearly with dose.

29.4.6 Conditioning on a Confounder

Underlying principle: X is a continuous confounder. The aim is to adjust for the confounding effect of X, but not to estimate its effect on the outcome variable.

There is no primary interest in assessing the relation between X and Y if X is a confounder of the variable Z where Z is the variable of primary interest. Here, the main focus lies on the correct and complete adjustment of the effect of Z for X. All previous methods may be applied; however, all these methods cannot exclude residual confounding. We therefore suggest to eliminate the effect of the confounder X by conditioning on X using fine strata. This method can readily be applied to case-control studies. It means that the regression coefficients of the logistic model are estimated by maximizing the conditional likelihood, where it is conditioned on the observations in the matched sets which are formed post hoc with fine strata according to X. Details may be found in Neuhäuser and Becher (1997). This method performs rather well in simulation studies with respect to bias and precision of the estimates in comparison to traditional methods to adjust for confounding by inclusion in the model as a covariate.

To further illustrate this idea, let us assume a case-control study with n_0 controls and n_1 cases. The logistic model with covariates X and Z reads as follows:

$$\text{logit } P(Y = 1|X = x) = \beta_0 + \beta_1 z + \beta_2 f(x).$$

The unconditional likelihood function of this logistic model is then given as

$$L_{\text{uncond}} = \prod_{i=1}^{n_0} \frac{1}{1 + \exp(\beta_0 + \beta' f(x) + \gamma z)} \prod_{i=n_0+1}^{n_0+n_1} \frac{\exp(\beta_0 + \beta' f(x) + \gamma z)}{1 + \exp(\beta_0 + \beta' f(x) + \gamma z)}.$$

To adjust for the confounding effect of X, it is common practice to include X into the model either untransformed (see Sect. 29.4.2), as a categorical variable (see Sect. 29.4.1), or using another transformations according to the methods described in Sect. 29.4.3. Instead, it is proposed to adjust for X via the conditional likelihood as

$$L_{cond} = \prod_{h=1}^{H} \frac{\prod_{j=1}^{n_{1h}} \exp(\gamma z_{hj})}{\sum_{l_j} \prod_{j=1}^{n_{1h}} \exp(\gamma z_{hl_j})}.$$

Here, we have defined H strata according to an appropriate categorization of X. Each stratum consists of n_{1h} cases and n_{0h} controls. The sum in the denominator ranges over the $l_j = \binom{n_{1h} + n_{0h}}{n_{1h}}$ choices to select n_{1h} observations out of the total number of observations in stratum h. It is possible to extend the method to more than one covariate, for example, sex and age. In that case, all individuals with the same sex and the same age group build one stratum. Since the likelihood is conditioned on the observations of the confounder X, this variable cancels out from the likelihood (Breslow and Day 1980).

This post hoc stratification method has a practical limitation. On the one hand, one aims at defining fine strata to avoid residual confounding. On the other hand, if these strata are too fine, several of these may contain no cases or no controls. For example, if strata are defined through "year and month of birth," then in a medium-sized study many will be uninformative since either cases or controls are missing. In that case, they do not contribute to the estimation since they cancel out from the likelihood function above. This, in turn, results in a loss of power. See Example 29.8 for a practical application of the method.

For an individually matched case-control study, this method can only be used to control for another confounder if the original matching is broken and new strata are formed according to the primary matching factors and the confounder in addition. Properties of this procedure have not yet been investigated.

Example 29.8.
We consider an unmatched case-control study on laryngeal cancer risk factors (Zatonski et al. 1991). Here, age is the continuous covariate X, and Z, smoking, is the variable of main interest. We are interested in an optimal adjustment for the confounding effect of X, not in a dose-effect analysis. Since the age distribution differs between cases and controls (see Table 29.7), adjustment for age is necessary.
Coding age as a dummy variable as described in Sect. 29.4.1 was used in the original paper with six age categories in an unconditional logistic regression model. However, this method may not be fully satisfactory as some residual confounding may still remain. Here, the proposed approach to condition on the confounder may be advisable. As alternative approach fractional polynomials could be used. Table 29.8 shows the estimated regression coefficients for cigarette consumption, given as log (no. of cigarettes smoked +1), and standard errors for different adjustment methods. Due to the confounding of the association of laryngeal cancer and smoking by age, the regression coefficients for smoking become smaller, the better the adjustment for age is. Considering age as dummy variable in the model (see Sect. 29.4.1) with six age categories reduces the estimate for smoking from

Table 29.7 Age distribution in cases and controls, Polish laryngeal cancer case-control study (Zatonski et al. 1991)

	Controls		Cases	
Age group (years)	N	%	N	%
25–34	275	28.5	7	2.8
35–39	163	16.9	7	2.8
40–44	96	10.0	19	7.6
45–49	90	9.3	27	10.9
50–54	110	11.4	71	28.5
55–59	123	12.7	62	24.9
60–65	108	11.2	56	22.5
Total	965	100.0	249	100.0

Table 29.8 Selected regression coefficients for smoking by adjustment method, Polish laryngeal cancer case-control study (Zatonski et al. 1991)

Method of adjustment for age	Deviance	$\hat{\beta}_{\text{smoking}}$	Standard error (s.e.)
No adjustment	960.4	1.732	0.153
Categorical (2 age groups)	946.0	1.440	0.162
Categorical (6 age groups)	919.7	1.236	0.163
Fractional polynomial degree 1 (age^{-2})	931.6	1.193	0.160
Fractional polynomial degree 2 (age^3, log(age)×age^3)	918.7	1.185	0.159
41 age strata (1-year interval)	–	1.161	0.159

1.732 to 1.236. The best fractional polynomials of degrees 1 and 2 yield an estimate of 1.193 and 1.185, respectively. Conditioning on age using 1-year intervals further reduces the estimate to 1.161. The differences between these latter estimates do not have practical relevance; however, they confirm the simulations by Neuhäuser and Becher (1997) showing that conditioning is an appropriate method to control for the confounding effect of X since (i) the confounding effect of age appears to be adjusted best and (ii) the standard error of the estimate is virtually identical to that of the other models.

29.4.7 Generalized Additive Models (GAM's)

Underlying principle: The effect of X on the outcome variable is estimated by an unspecified function.

The class of generalized additive models (GAM's) proposed by Hastie and Tibshirani (1986, 1990) is a method to model an unspecified relation between a set of k covariates $X_1, \ldots, X_k$ and a response variable Y.

In the framework of the logistic regression model and for the simplest case with one covariate X, the common logistic model logit $P(Y = 1|x) = \beta_0 + \beta_1 x$ is replaced by logit $P(Y = 1|x) = \beta_0 + f(x)$ or logit $P(Y = 1|x) = \beta_0 + \beta_1 x + f(x)$ where $f(x)$ is a unspecified ("non-parametric") function. This function is estimated in a flexible manner using a scatterplot smoother. The estimated function $\hat{f}(x)$ can reveal possible non-linearities in the effect of X. Such unspecified smooth functions are estimated by local scoring algorithms further described in Hastie and Tibshirani (1986, 1990). This method has been used in several epidemiological studies on air pollution and health effects (e.g., Stieb et al. 2000; Rossi et al. 1999). The method is intuitively appealing since it provides a flexible method for identifying non-linear covariate effects. However, this approach is more data driven than the previous techniques described in this chapter, and it is not possible to take biological knowledge of the shape of the dose-effect curve into account. In some software packages, for example, in SAS, a procedure is available (PROC GAM) to fit a generalized additive model. It has the disadvantage, however, that it is not possible to plot the dose-effect curve (see Example 29.9).

Example 29.9.
In this example, we illustrate the application of the GAM and compare it with the standard logistic regression approach. We use the dataset from the laryngeal cancer study and estimate the risk associated with smoking among smokers (non-smokers are excluded from the analysis). We first fit the common logistic model logit $P(Y = 1|x) = \beta_0 + \beta_1 x$ with X being the cumulative amount of smoking (in pack-years/10). Using the procedure LOGISTIC in SAS using the statement

```
Proc logistic descending; model y=x;
```

the logistic model can be estimated, and we get a highly significant estimate for the effect of smoking on laryngeal cancer ($\hat{\beta}_1 = 0.36, \text{s.e.}(\hat{\beta}_1) = 0.037, p < 0.0001$). To check whether the log odds ratio increases linearly with dose, we fit a GAM using the statement

```
Proc gam plots=components(clm commonaxes);model y
(event='1') = spline(x,df=3)/dist=binomial;
```

Using this specification, the procedure estimates a linear component β_1 and a smoothing function $f(x)$ to describe the deviation from linearity over the observation range. Here, we get a highly significant linear component ($\hat{\beta}_1 = 0.33, \text{s.e.}(\hat{\beta}_1) = 0.037, p < 0.0001$) and also a highly significant smoothing component ($\chi^2 = 24.4, p < 0.0001$). The plot of the smoothing component (Fig. 29.2) suggests that the risk increase is strongest in the range of two to six and lower otherwise.
When including $\log(x)$ in the model instead of x, we get a highly significant linear component ($\hat{\beta}_1 = 1.03, \text{s.e.}(\hat{\beta}_1) = 0.11, p < 0.0001$), whereas the smoothing component ($\chi^2 = 2.65, p = 0.27$) is no longer significant. The corresponding plot of the smoothing component (Fig. 29.3) shows that the smoothing component does indeed differ only slightly from zero.

29.4.8 Spline Regression

Underlying principle: Spline regression is a smoothly joined piecewise regression. It is a non-parametric regression technique and can be seen as an extension of linear models that automatically models non-linearities and interactions between variables with a large flexibility:

$$f(x) \text{ polynomial (see below); } R(\cdot) = \exp(\cdot).$$

Greenland (1995) has proposed spline regression within the logistic model (see also chapter ▶Regression Methods for Epidemiological Analysis of this handbook) as a valuable method when important non-linearities are anticipated and software for more general non-parametric regression approaches is not available. Desquilbet and Mariotti (2010) described restricted cubic splines for dose-response analyses and present a SAS macro for analysis. Boucher et al. (1998) used this method for analyzing a case-control study on colon cancer and described several advantages of the method in comparison to the categorical analysis based on dummy variables (see Sect. 29.4.1). While using dummies, the risk is assumed to be constant within a

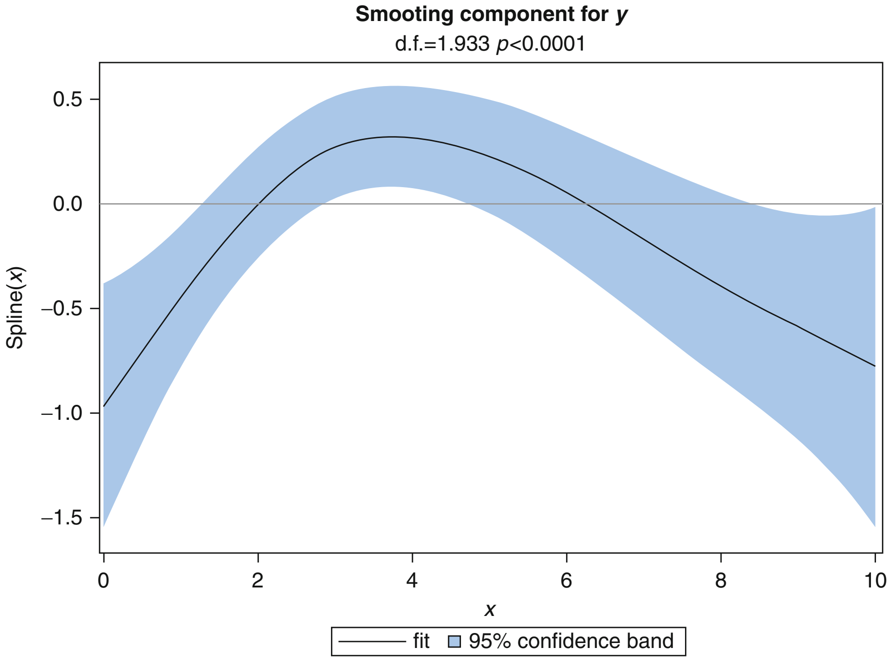


Fig. 29.2 Plot of the smoothing component

category and increases stepwise from category to category; with spline regression, the fitted risk changes continuously within and across categories. The method can also be applied to the other regression models considered in this chapter. Spline regression is formally described as follows: Let X be divided into K categories indexed by $k = 1, \ldots, K$ with $K - 1$ internal cut-points $c_1, \ldots, c_{K-1}$. For the so-called linear spline regression, we have $f(x) = \alpha + \beta_1 x + \beta_2 s_2 + \ldots + \beta_K s_K$, where $s_k = 0$ if $x \leq c_k$ and $s_k = x - c_k$ if $x > c_k$ and $R(\cdot) = \exp(\cdot)$. Using this model, one gets a continuous dose-effect function with slopes changing at each cut-point.

A more general spline regression avoids the biologically implausible fact that the first derivative of the dose-effect function is not continuous. Here, a quadratic term for X is added, and squared values of s_k are entered into the model such that

$$f(x) = \alpha + \beta_1 x + \gamma_1 x^2 + \beta_2 s_2^2 + \ldots + \beta_K s_K^2.$$

More details can be found in Greenland (1995) and Desquilbet and Mariotti (2010).

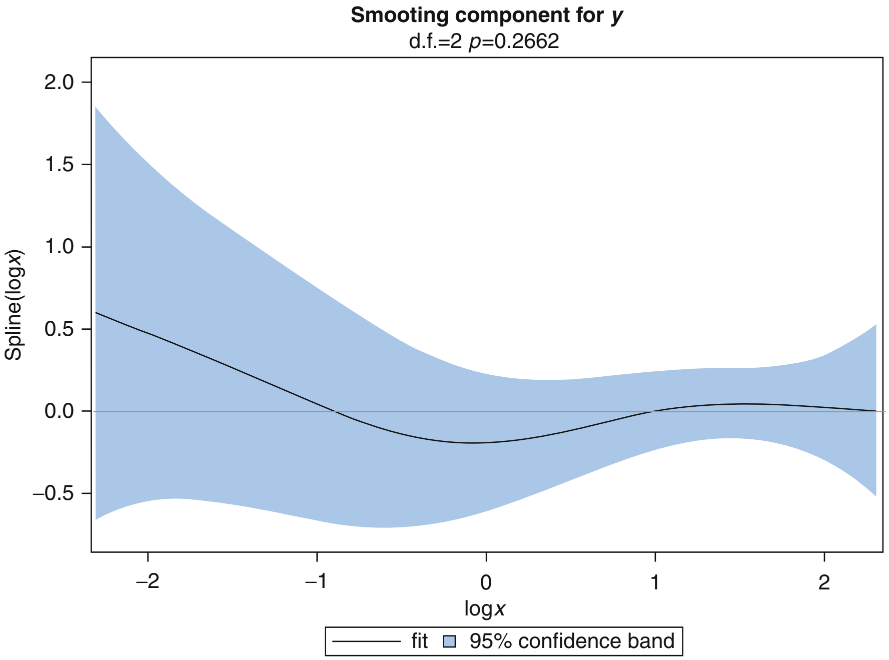


Fig. 29.3 Plot of the smoothing component

29.5 Conclusions

After a descriptive analysis, which may include some of the methods described in Sect. 29.2 – some of these are automatically carried out by some software packages when contingency tables are generated – more elaborate methods come into play. But there is no fixed analysis strategy such that each possible situation is treated appropriately. However, some general recommendations may help to develop an appropriate strategy in a given situation.

Figure 29.4 provides a possible decision tree which is further described below.

If X is not the variable of primary interest, conditioning is the recommended method in case-control studies to control for the confounding effect of X on the covariate of interest to minimize residual confounding. Some care must be taken to define the strata appropriately. For example, if an adjustment for age is done by defining strata as "year of birth," this will result in large strata for the most frequent age groups and small strata for the extreme ages. In Neuhäuser and Becher (1997), an adaptive method which depends on the study size and distribution of X is further discussed in the framework of the logistic regression. For other regression models, this method for control of confounding has not yet been investigated.

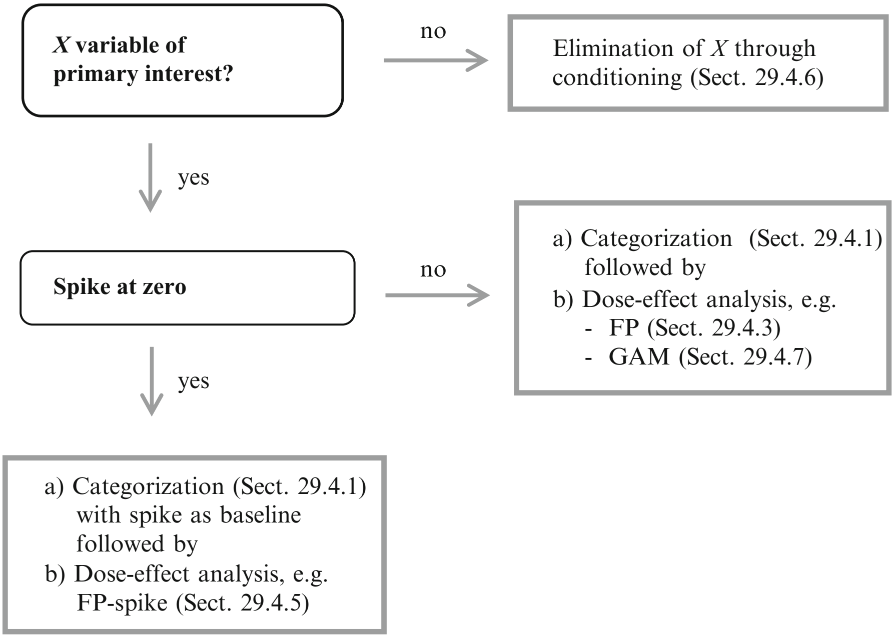

X variable of primary interest?
no
Elimination of *X* through conditioning (Sect. 29.4.6)
yes
Spike at zero
no
a) Categorization (Sect. 29.4.1) followed by
b) Dose-effect analysis, e.g.
- FP (Sect. 29.4.3)
- GAM (Sect. 29.4.7)
yes
a) Categorization (Sect. 29.4.1) with spike as baseline followed by
b) Dose-effect analysis, e.g. FP-spike (Sect. 29.4.5)

Fig. 29.4 Suggested decision tree for including continuous covariates

If there are several confounders simultaneously for which the risk estimate is not of interest, the method can in principle be used by forming strata according to the cross-classification of all these variables. If, however, the number of strata becomes large in comparison to the number of observations, the method has reached its limit.

If X is the variable of primary interest, the situation is more complex. If possible, biological knowledge should be taken into account when choosing the regression model. If the risk function is assumed to be non-monotonic, fractional polynomials are a good choice. Another option would be GAM's or spline regressions; however, as Royston et al. (2000) pointed out, the latter may exhibit artifacts which can make their interpretation difficult. But if the risk function is monotonic, FPs can be used for modeling the relationship between the continuous covariate X and the outcome by simply restricting to FPs of degree 1.

For the less frequent case that not the dose-effect curve itself is of interest but simply the hypothesis that there is a relation between the variable X and the response, the recommended procedure is as follows: Perform the analysis for X as for estimating the dose-effect curve and use the corresponding p-value.

Many of the methods presented above for assessing the dose-effect curve can yield very similar results although they may look very different at first sight. The models are often very closely related, and obviously the power to detect model misspecification is very small. While a categorization into two categories (high/low exposure) is generally not recommended, a categorization into four or more categories is, although not the optimal method, a useful first step in course of the analysis because it gives a first impression on the effect of the variable for different dose levels. This is especially true if a natural baseline level exists as for smoking where this is the case for non-smokers. For nutritional variables like "fat consumption," it is more difficult to define an appropriate baseline. Among the common methods are the construction of tertiles, quartiles, etc., or to use the same cut-points as in previous studies in order to allow a comparison of the results.

The availability of appropriate software is an important practical aspect since typically it is not possible to develop new software tools or even new statistical methods when data are routinely analyzed within the framework of an epidemiological study.

Both GAM's and spline regression have drawbacks; however, user-friendly software is available, and as an exploratory tool, these may be useful. As Royston et al. (2000) pointed out, "there are artefacts in the curve shapes, and possible overinterpretation of the data that may accompany them." Royston et al. (1999) stated, "we do not think that [GAM's and spline regression] are suitable as definitive models in epidemiology for the following reasons. The mathematical expressions for the curves are often very complex, so reporting of results must be by graphs or by extensive tabulation. The situation is unsatisfactory when similar studies are to be compared, and impossible if meta-analysis is intended. Data dependence of the final model is more marked than for parametric models and the curves may be more difficult to interpret." We also think that the suggested parametric models such as the fractional polynomial method with or without spike, depending on the situation, provide sufficient flexibility on the one hand and allow inclusion of a priori assumptions on the shape of the dose-response curve (monotonous, U-shaped) on the other hand.

If a complex epidemiological study is analyzed independently by different statisticians, the results will typically not be identical. This is particularly true when continuous covariates are among the variables to be considered (not to mention other issues that make the analysis of epidemiological studies non-standard, such as the treatment of missing values, variable selection procedures, model building, and others). Readers can be reassured, however, that if an analysis is performed with sufficient care, the results and their interpretation will not differ very much.

Acknowledgements Effective help by Eva Lorenz with the calculations of the examples is gratefully acknowledged.

References

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, Hoboken

Altman DG, Lausen B, Sauerbrei W, Schumacher M (1994) Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst 86:829–835

Becher H (1993) The concept of residual confounding in regression models and some applications. Stat Med 11:1747–1758

Becher H, Lorenz E, Royston P, Sauerbrei W (2012) Analysing covariates with spike at zero: a modified FP procedure and conceptual issues. Biom J 54:686–700

Bernstein S, Bernstein R (1998) Schaum's outline of elements of statistics I: descriptive statistics and probability. McGraw-Hill, New York

Boucher KM, Slattery ML, Berry TD, Quesenberry C, Anderson K (1998) Statistical methods in epidemiology: a comparison of statistical methods to analyze dose-response and trend analysis in epidemiologic studies. J Clin Epidemiol 51:1223–1233

Breslow N, Day N (1980) Statistical methods in cancer research. Volume I – the analysis of case-control studies. IARC Scientific Publications No. 32. International Agency for Research on Cancer, Lyon

Breslow N, Day N (1987) Statistical methods in cancer research. Volume II – the design and analysis of cohort studies. IARC Scientific Publications No. 82. International Agency for Research on Cancer, Lyon

Breslow NW, Storer BE (1985) General relative risk functions for case-control studies. Am J Epidemiol 122:149–162

Desquilbet L, Mariotti F (2010) Dose-response analyses using restricted cubic spline functions in public health research. Stat Med 29:1037–1057

Greenland S (1995) Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology 6:356–365

Hastie T, Tibshirani R (1986) Generalized additive models. Stat Sci 1:297–318

Hastie T, Tibshirani R (1990) Generalized additive models. Chapman & Hall, London

Jedrychowski W, Becher H, Wahrendorf J, Basa-Cierpialek Z, Gomola G (1992) Effect of tobacco smoking on various histologic types of lung cancer. J Cancer Res Clin Oncol 118: 276–282

Maclure M, Greenland S (1992) Tests for trend and dose response: misinterpretations and alternatives. Am J Epidemiol 135:96–104

Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 22:719–748

Mizon GE, Richard JF (1986) The encompassing principle and its application to testing non-nested hypothesis. Econometrica 54:657–678

Neuhäuser M, Becher H (1997) Improved odds ratio estimation by posthoc stratification of case-control data. Stat Med 16:993–1004

Olsen MK, Schafer JL (2001) A two-part random-effects model for semicontinuous longitudinal data. J Am Stat Assoc 96:730–745

Robertson C, Boyle P, Hsieh CC, Macfarlane GJ, Maisonneuve P (1994) Some statistical considerations in the analysis of case-control studies when the exposure variables are continuous measurements. Epidemiology 5:164–170

Rossi G, Vigotti MA, Zanobetti A, Repetto F, Gianelle V, Schwartz J (1999) Air pollution and cause-specific mortality in Milan, Italy, 1980–1989. Arch Environ Health 54:158–164

Royston P, Altman DG (1994) Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Appl Stat 43:429–467

Royston P, Sauerbrei W (2008) Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Wiley, Chichester

Royston P, Thompson SG (1995) Comparing non-nested regression models. Biometrics 51: 114–127

Royston P, Ambler G, Sauerbrei W (1999) The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol 28:964–974

Royston P, Sauerbrei W, Altman DG (2000) Modeling the effects of continuous risk factors [letter]. J Clin Epidemiol 53:219–221

Royston P, Sauerbrei W, Becher H (2010) Modelling continuous exposures with a 'spike' at zero: a new procedure based on fractional polynomials. Stat Med 29:1219–1227

Schisterman EF, Reiser B, Faraggi D (2006) ROC analysis for markers with mass at zero. Stat Med 25:623–638

Schulgen G, Lausen B, Olsen JH, Schumacher M (1994) Outcome-oriented cutpoints in analysis of quantitative exposures. Am J Epidemiol 140:172–184

Stieb DM, Beveridge RC, Brook JR, Smith-Doiron M, Burnett RT, Dales RE, Beaulieu S, Judek S, Mamedov A (2000) Air pollution, aeroallergens and cardiorespiratory emergency department visits in Saint John, Canada. J Expo Anal Environ Epidemiol 10:461–477

Zatonski W, Becher H, Lissowska J, Wahrendorf J (1991) Tobacco, alcohol and diet in the etiology of laryngeal cancer – a population-based case-control study. Cancer Causes Control 2:3–10

Regression Methods for Epidemiological Analysis 30

Sander Greenland

Contents

S. Greenland
Department of Epidemiology, School of Public Health, University of California, Los Angeles, CA, USA

Department of Statistics, UCLA College of Letters and Science, Los Angeles, CA, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_17,

30.1 Introduction

Basic tabular and graphical methods are an essential component of epidemiological analysis and are often sufficient, especially when one need consider only a few variables at a time. They are, however, limited in the number of variables that they can examine simultaneously and in detail they can consider continuous variables. Even sparse-strata methods (such as Mantel-Haenszel) require that some strata have two or more subjects; yet, as more and more variables or categories are added to a stratification, the number of subjects in each stratum may eventually drop to 0 or 1.

Regression analysis encompasses a vast array of techniques designed to overcome the numerical limitations of simpler methods. This advantage is purchased at a cost of stronger assumptions, which are compactly represented by a *regression model.* Such models (and hence the assumptions they represent) have the advantage of being explicit; a disadvantage is that the models may not be well understood by the intended audience or even the user. Regression models can and should be tailored by the analyst to suit the topic at hand; the latter process is sometimes called *model specification.* This process is part of the broader task of *regression modeling*.

To ensure that the assumptions underlying the regression analysis are reasonable approximations, it is essential that the modeling process be actively guided by the scientists involved in the research, rather than be left solely to mechanical algorithms. Such active guidance requires familiarity with the variety and interpretation of models. Hence, this chapter will focus primarily on forms of models and their interpretation, rather than on the more technical issues of model fitting and testing. Because this chapter provides only outlines of key topics, it should be supplemented by readings in more detailed treatments of regression analysis, as can be found in texts such as McCullagh and Nelder (1989), Hosmer and Lemeshow (2000), and Jewell (2004). For a classic in-depth treatment of the difficulties and limitations of regression analysis in non-experimental studies, see Leamer (1978) or Berk (2004).

Achieving competence in regression analysis requires comfort with basic geometry and algebra. While the ensuing discussion attempts to be self-contained, readers who feel lacking or weak in mathematical skills would do well to review a textbook in high school mathematics or college algebra (focusing especially on functions, graphs, and natural logarithms) before studying regression methods.

30.2 Regression Functions

There are two primary interpretations of regression, frequentist and Bayesian, which correspond to two different interpretations of probability (see Greenland 2006 for a basic comparison of interpretations). This chapter uses the frequentist interpretation, but briefly discusses the Bayesian interpretation at the end of this section. In both interpretations, the term *regression* is often used as a shorthand for *regression function*.

30.2.1 Frequentist Regression

In the frequentist view, the *regression* of a variable Y on another variable X is the function that describes how the average (mean) value of Y changes across population subgroups defined by levels of X (sometimes the median or some other quantile is used in place of the mean). This function is distinct from a *model* for that function. A regression *model* is another, simpler function used to approximate or estimate the true regression function. This distinction is often obscured and even unrecognized in elementary treatments of regression, which in turn has generated much misunderstanding of regression modeling. Therefore, this chapter will begin by focusing on regression functions and then will turn to regression models.

The regression function is often written as $E(Y|X = x)$, which should be read as "the average of Y when the variable X takes on the specific value x." The "E" part of the notation stands for "expectation," which here is just another word for "population mean." As an example, suppose Y stands for "height" to the nearest centimeter at some time t, X stands for "weight" to the nearest kilogram at time t, and the population of interest is that of Denmark at time t. If we subclassify the Danish population at t into categories of weight X, compute the average height in each category and tabulate or graph these average heights against the weight categories, the result displays the regression, $E(Y|X = x)$, of height Y on weight X in Denmark at time t. Several important points should be emphasized:

1. The *concept* of regression involves no modeling. Some would describe this fact by saying that the concept of regression is essentially "non-parametric." The regression of Y on X is just a graphical property of the physical world, like the orbital path of the earth around the sun.
2. There is nothing mathematically sophisticated about the regression function. Each point on a regression curve could be computed by taking the average of Y within a subpopulation defined as having a particular value of X. In the example, the value of the regression function at $X = 50$ kg, $E(Y|X = 50)$, is just average height at time t among Danes who weigh 50 kg at time t.
3. A regression function cannot be unambiguously computed until we carefully define X, Y, *and* the population over which the averages are to be taken. We will call the latter population the *target population* of the regression. This population is all too often left out of regression definitions, often resulting in confusion.

   Some ambiguity is unavoidable in practice. In our example, is time t measured to the nearest year, day, minute, or millisecond? Is the Danish population all citizens, all residents, or all persons present in Denmark at t? We may decide that leaving these questions unanswered is tolerable because varying the definitions over a modest range would not change the result to an important extent. But if we left time completely out of the definition, the regression would become hopelessly ambiguous, for now we would not have a good idea of who to include or exclude from our average: Should we include people living in Denmark in prehistoric times or in the time of King Canute (a 1,000 years ago) or in the distant future (a 1,000 years from now)? The choice could have a strong effect on our answer because of the large changes in height-to-weight relations that have occurred over time.

When considering a regression function $E(Y|X = x)$, the variable Y is termed the dependent variable, outcome variable, or *regressand,* and the variable X is termed the independent variable, predictor, covariate, or *regressor.* The "dependent/independent" terminology is common but also problematic because it invites confusion of distinct probabilistic and causal concepts of dependence and independence. For example, if Y is age and X is blood pressure, $E(Y|X = x)$ represents the average age of persons given blood pressure, X. But it is blood pressure X that causally depends on age Y, not the other way around. This brings us to another important point.

4. For any pair of variables X and Y, we can consider either the regression of Y on X, $E(Y|X = x)$, or the regression of X on Y, $E(X|Y = y)$. Thus, the concept of regression does not necessarily imply any causal or even temporal relation between the regressor and the regressand.

For example, Y could be blood pressure at the start of follow-up of a cohort, and X could be blood pressure after 1 year of follow-up; then $E(Y|X = x)$ would represent the average *initial* blood pressure among cohort members whose blood pressure after 1 year of follow-up is x. This is an example of a non-causal regression. Causal regression will be discussed below.

30.2.2 Other Concepts of Population

It is important to distinguish between a "target population" and a "source population." The target population of regression is defined without regard to our observations; for example, the regression of diastolic blood pressure on cigarette usage in China is defined whether or not we conduct a study in China (the target for this regression). A source population is a source of subjects for a particular study and is defined by the selection methods of the study; for example, a random-sample survey of all residents of Beijing would have Beijing as its source population. The concepts of target and source populations connect only insofar as inferences about a regression function drawn from a study are most easily justified when the source population of the study is identical to the target population of the regression. Otherwise, issues of generalization from the source to the target have to be addressed (see Rothman et al. 2008, Chap. 9).

In some literature, regression functions (and many other concepts) are defined in terms of averages within a "superpopulation" or "hypothetical universe." A superpopulation is an abstraction of a target population, sometimes said to represent the distribution (with respect to all variables of interest) of all possible persons that ever were or ever could be targets of inference for the analysis at hand. Because the superpopulation approach focuses on purely hypothetical distributions, it has encouraged substitution of mathematical theory for the more prosaic task of connecting study results to populations of immediate public-health concern. Thus, this chapter defines regression functions in terms of averages within real (target) populations.

30.2.3 Binary Regression

The concept of regression applies to variables measured on any scale: The regressand and the regressor may be continuous or discrete or even binary. For example, Y could be an indicator of diabetes ($Y = 1$ for present, $Y = 0$ for absent), and X could be an indicator for sex ($X = 1$ for female, $X = 0$ for male). Then $E(Y|X = 1)$ would represent the average of the diabetes indicator Y among females, and $E(Y|X = 0)$ would represent the average of Y among males.

When the regressand Y is a binary indicator (0, 1) variable, $E(Y|X = x)$ is called a *binary regression,* and this regression simplifies in a very useful manner. Specifically, when Y can be only 0 or 1, the average $E(Y|X = x)$ equals the proportion of population members who have $Y = 1$ among those who have $X = x$. For example, if Y is the diabetes indicator, $E(Y|X = x)$ is the proportion with diabetes (i.e., with $Y = 1$) among those with $X = x$. To see this, let N_{yx} denote the number of population members who have $Y = y$ and $X = x$. Then the number of population members with $X = x$ is $N_{1\mathrm{x}} + N_{0\mathrm{x}} = N_{+\mathrm{x}}$, and the average of Y among these members, $E(Y|X = x)$, is

$$\frac{N_{1x} \cdot 1 + N_{0x} \cdot 0}{N_{1x} + N_{0x}} = \frac{N_{1x}}{N_{+x}},$$

which is just the proportion with $Y = 1$ among those with $X = x$.

The epidemiological ramifications of the preceding relation are important. Let $\Pr(Y = y|X = x)$ stand for "the proportion (of population members) with $Y = y$ among those with $X = x$" (which is often interpreted as the probability of $Y = y$ in the subpopulation with $X = x$). If Y is a binary indicator, we have just seen that

$$E(Y|X = x) = \Pr(Y = l|X = x),$$

that is, the average of Y when $X = x$ equals the proportion with $Y = 1$ when $X = x$. Thus, if Y is an indicator of *disease presence* at a given time, the regression of Y on X, $E(Y|X = x)$, provides the proportion *with* the disease at that time, or prevalence proportion, given $X = x$. For example, if $Y = 1$ indicates diabetes presence on January 1, 2010, and X is weight on that day, $E(Y|X = x)$ provides diabetes prevalence as a function of weight on that day. If Y is instead an indicator of *disease incidence* over a time interval (cf. Rothman et al. 2008, Chap. 3, and chapter ►Rates, Risks, Measures of Association and Impact of this handbook), the regression of Y on X provides the proportion getting disease over that interval, or incidence proportion, given $X = x$. For example, if $Y = 1$ indicates stroke occurrence in 2010 and X is weight at the start of the year, $E(Y|X = x)$ provides the stroke incidence (proportion) in 2010 as a function of initial weight.

30.2.4 Multiple Regression

The concept of multiple regression is a simple extension of the ideas discussed above to situations in which there are multiple (two or more) regressors. To illustrate, suppose Y is a diabetes indicator, X_1 stands for "sex" (coded 1 for females, 0 for males), and X_2 stands for "weight" (in kilograms). Then the regression of Y on X_1 and X_2, written $E(Y|X_1 = x_1, X_2 = x_2)$, provides the average of Y among population members of a given sex X_1 and weight X_2. For example, $E(Y|X_1 = 1, X_2 = 70)$ is the average diabetes indicator (and, hence, the diabetes prevalence) among women who weigh 70 kg.

We can use as many regressors as we want. For example, we could include age (in years) in the last regression. Let X_3 stand for "age." Then $E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$ would provide the diabetes prevalence among population members of a given sex, weight, and age. Continuing to include regressors produces a very clumsy notation, however, and so we adopt a simple convention: We will let $\boldsymbol{X}$ represent the ordered list of all the regressors we want to consider. Thus, in our diabetes example, $\boldsymbol{X}$ will stand for the horizontal list (X_1, X_2, X_3) of "sex," "weight," and "age." Similarly, we will let $\boldsymbol{x}$ stand for the horizontal ordered list of values (x_1, x_2, x_3) for $\boldsymbol{X} = (X_1, X_2, X_3)$. Thus, if we write $E(Y|\boldsymbol{X} = \boldsymbol{x})$, it is merely a shorthand for

$$E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

when there are three regressors under consideration.

More generally, if there are n regressors $X_1, \ldots, X_n$, we will write $\boldsymbol{X}$ for the ordered list $(X_1, \ldots, X_n)$ and $\boldsymbol{x}$ for the ordered list of values $(x_1, \ldots, x_n)$. The horizontal ordered list of variables X is called a *row vector* of regressors, and the horizontal ordered list of values $\boldsymbol{x}$ is called a *row vector* of values. Above, the vector $\boldsymbol{X}$ is composed of the $n = 3$ items "sex," "weight," and "age," and the list $\boldsymbol{x}$ is composed of specific values for sex (0 or 1), weight (kilograms), and age (years). The number of items n in $\boldsymbol{X}$ is called the length or dimension of $\boldsymbol{X}$.

The term *multivariate* (or multiresponse) *regression* is usually reserved for regressions in which there are multiple *regressands* (Izenman 2008). To illustrate, suppose Y_1 is an indicator of diabetes presence, Y_2 is diastolic blood pressure, and $\boldsymbol{Y}$ is the list (Y_1, Y_2) composed of these two variables. Also, let $\boldsymbol{X}$ be the list (X_1, X_2, X_3) composed of the sex indicator, weight, and age, as before. The multivariate regression of diabetes and blood pressure on sex, weight, and age provides the average diabetes indicator *and* average blood pressure for each combination of sex, weight, and age:

$$E(Y_1, Y_2|X_1 = x_1, X_2 = x_2, X_3 = x_3) = E(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}).$$

In general, there may be any number of regressands in the list $\boldsymbol{Y}$ and regressors in the list $\boldsymbol{X}$ of a multivariate regression. Multivariate regression notation allows one to express the separate regressions for each regressand in one equation.

30.2.5 Regression and Causation

Because regression functions do not involve any assumptions of time order or causal relations, regression coefficients and quantities derived from them represent measures of association, not measures of effect. To interpret the coefficients as measures of causal effects, it is important that the regression function being modeled provides a representation of the effects of interest that is approximately unconfounded (for a general discussion of the concept of confounding, see Rothman et al. 2008, Chap. 4 and chapter ▶Confounding and Interaction of this handbook).

To make this no-confounding assumption more precise, suppose $\boldsymbol{X}$ contains the exposures of interest and $\boldsymbol{Z}$ contains the other regressors. Suppose we write "Set($\boldsymbol{X}=\boldsymbol{x}$)" for the act of setting (forcing) the value of $\boldsymbol{X}$ to $\boldsymbol{x}$, assuming that can be done; another notation for the same concept is "Do($\boldsymbol{X}=\boldsymbol{x}$)" (Pearl 2009). We may then write

$$E[Y|\text{Set}(\boldsymbol{X}=\boldsymbol{x}),\ \boldsymbol{Z}=z]$$

for the average value Y would have *if* everyone in the target population with $\boldsymbol{Z}=z$ had their $\boldsymbol{X}$ value set to $\boldsymbol{x}$. In other words, it is what the mean of Y would have been among those with $\boldsymbol{Z}=z$ if everyone in the population had received $\boldsymbol{X}=\boldsymbol{x}$ (which is what "Set" means).

There are many variables (such as sex and age) for which the use of "Set" will make little sense or be too vague to be useful (Greenland 2005a; Hernán 2005); such variables should thus appear only in the stratification variables ($\boldsymbol{Z}$ in the above expression). But even for potential intervention variables for which "Set" can makes sense (such as smoking behavior or air-pollutant levels), an average like $E[Y|\text{Set}(\boldsymbol{X}=\boldsymbol{x}),\ \boldsymbol{Z}=z]$ cannot be estimated from observational data without making strong assumptions because it may represent a *counterfactual* state, that is, a condition that did not in fact occur. In particular, it can be very different from the observable average $E(Y|\boldsymbol{X}=\boldsymbol{x},\ \boldsymbol{Z}=z)$. The latter refers only to those population members with $\boldsymbol{X}=\boldsymbol{x}$ and $\boldsymbol{Z}=z$, whereas $E[Y|\text{Set}(\boldsymbol{X}=\boldsymbol{x}),\ \boldsymbol{Z}=z]$ refers to *all* population members with $\boldsymbol{Z}=z$, including those who actually had $\boldsymbol{X}$ equal to values other than $\boldsymbol{x}$.

As an example, suppose the target population is all persons born during 1901–1950 surviving to age 50, Y is an indicator of death by age 80, $\boldsymbol{X}$ contains only $X_1=$pack-years of cigarettes smoked by age 50, and $\boldsymbol{Z}=(Z_1,\ Z_0)$ where $Z_1=1$ if female, 0 if male and $Z_2=$year of birth. Then

$$E[Y|X_1=20,\ \boldsymbol{Z}=(1,1940)]$$

would be the average risk of dying by age 80 (mortality proportion) among women born in 1940 and surviving to age 50 who smoked 20 pack-years by age 50. In contrast,

$$E[Y|\text{Set}(X_1=20),\ \boldsymbol{Z}=(1,1940)]$$

would be the average risk of dying by age 80 among all women born in 1940 and surviving to age 50 *if* all such women had smoked 20 pack-years by age 50. If some women in this cohort smoked 10 pack-years or 30 pack-years or none by age 50, the latter average would be counterfactual and hence could be estimated only by making special assumptions that we now explore.

In regression analysis, we may define effect measures as contrasts of average outcomes (such as incidence) in the same population under different conditions. Consider the ratio effect measure contrasting the average of Y in the subpopulation with $\boldsymbol{Z} = \boldsymbol{z}$ when $\boldsymbol{X}$ is set to $\boldsymbol{x}^*$ versus that average when $\boldsymbol{X}$ is set to $\boldsymbol{x}$:

$$\frac{E[Y|\text{Set}(\boldsymbol{X} = \boldsymbol{x}^*), \boldsymbol{Z} = \boldsymbol{z}]}{E[Y|\text{Set}(\boldsymbol{X} = \boldsymbol{x}), \boldsymbol{Z} = \boldsymbol{z}]}.$$

In the example,

$$\frac{E[Y|\text{Set}(X_1 = 20),\ \boldsymbol{Z} = (1, 1940)]}{E[Y|\text{Set}(X_1 = 0),\ \boldsymbol{Z} = (1, 1940)]}$$

represents the *effect* of smoking 20 pack-years by age 50 versus no smoking on the risk of dying by age 80 among women born in 1940. On the other hand, the ratio measure

$$\frac{E[Y|X_1 = 20,\ \boldsymbol{Z} = (1, 1940)]}{E[Y|X_1 = 0,\ \boldsymbol{Z} = (1, 1940)]}$$

represents only the *association* of smoking 20 pack-years by age 50 versus no smoking with the risk among women born in 1940 because it contrasts two different subpopulations (one with $X_1 = 20$, the other with $X_1 = 0$).

To infer that all associational measures estimated from our analysis equal their corresponding effect measures, we would have to make the following assumption of no confounding given $\boldsymbol{Z}$ (which is sometimes expressed by stating that there is no residual confounding):

$$E(Y|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}) = E[Y|\text{Set}(\boldsymbol{X} = \boldsymbol{x}),\ \boldsymbol{Z} = \boldsymbol{z}].$$

This assumption states that the average we observe or estimate in the subpopulation with both $\boldsymbol{X} = \boldsymbol{x}$ and $\boldsymbol{Z} = \boldsymbol{z}$ is equal to what the average in the larger subpopulation with $\boldsymbol{Z} = \boldsymbol{z}$ would have been if everyone had $\boldsymbol{X}$ set to $\boldsymbol{x}$. It is important to appreciate the strength of the assumption. In the above example, the no-confounding assumption would entail

$$E[Y|X_1 = 20,\ \boldsymbol{Z} = (1, 1940)] = E[Y|\text{Set}(X_1 = 20),\ \boldsymbol{Z} = (1, 1940)].$$

This equation states that, in the cohort of women born in 1940 and surviving to age 50, the risk we observe among those who smoked 20 pack-years by age 50 equals the risk we would have observed if *all* women in the cohort had smoked 20 pack-years by age 50. The social variables associated with both smoking and death should lead us to doubt that the two quantities are even approximately equal.

If only a single summary measure of effect is desired, the covariate-specific no-confounding assumption can be replaced by a less restrictive assumption tailored to that measure. To illustrate, suppose in the above example, we are only interested in what the effect of smoking 20 versus 0 pack-years would be on *everyone* in the target, regardless of sex or birth year, as measured by the causal risk ratio

$$E[Y|\text{Set}(X_1 = 20)]/E[Y|\text{Set}(X_1 = 0)].$$

The corresponding measure of association is the risk ratio for 20 versus 0 pack-years, standardized to the total population:

$$\frac{\sum_z E(Y|X_1 = 20,\ \mathbf{Z} = z)\,\Pr(\mathbf{Z} = z)}{\sum_z E(Y|X_1 = 0,\ \mathbf{Z} = z)\,\Pr(\mathbf{Z} = z)},$$

where $\Pr(\mathbf{Z} = z)$ is the proportion with $\mathbf{Z} = z$ in the target. The no-confounding assumption we need here is that the standardized ratio equals the causal ratio. This summary assumption could hold even if there was confounding within levels of sex and birth year (although it would still be implausible in this example).

The dubiousness of no-confounding assumptions is often the chief limitation in using epidemiological data for causal inference. This limitation applies to both tabular and regression methods. Randomization of persons to levels of $\mathbf{X}$ can largely overcome this limitation because it ensures that effect estimates follow a quantifiable probability distribution centered around the true effect. Randomization is not an option in most settings, however.

The default strategy is to ensure there are enough well-measured confounders in $\mathbf{Z}$ so that the no-confounding assumption is at least plausible. This strategy often leads to few subjects at each level $\mathbf{x}$ of $\mathbf{X}$ and $\mathbf{z}$ of $\mathbf{Z}$, which in turn lead to the sparse-data problems that regression modeling attempts to address (Robins and Greenland 1986; Greenland 2000a, b; Greenland et al. 2000). A major limitation of this strategy is that, often, key confounders are poorly measured or unmeasured and so cannot be used in ordinary modeling; prior distributions for the missing confounders must be used instead (Greenland 2003, 2005b, 2009b; Greenland and Lash 2008; see also the chapter on Sensitivity Analysis and Bias Analysis in this handbook).

30.2.6 Frequentist Versus Bayesian Regression

In frequentist theory, an expectation is interpreted as an average in a specific subgroup of a specific population. The regression $E(Y|\mathbf{X} = \mathbf{x})$ thus represents an objective functional relation among theoretically measurable variables (the average of Y as a function of the variables listed in $\mathbf{X}$). It may be that this relation has not been observed, perhaps because it exists but we are unable to measure it or because it does not yet exist. Examples of the former and latter are the regressions of blood pressure on weight in Spain 10 years ago and 10 years from now. In either situation, the regression is an external relation that one tries to estimate, perhaps by projecting (extrapolating) from current knowledge about presumably similar relations. For example, one might use whatever survey data one can find on blood pressure and weight to estimate what the regression of blood pressure on weight would look like in Spain 10 years ago or 10 years from now. In this approach, one tries to produce an *estimate* $\hat{E}(Y|\boldsymbol{X} = \boldsymbol{x})$ of the true regression $E(Y|\boldsymbol{X} = \boldsymbol{x})$.

In subjective Bayesian theory (cf. Greenland 2006 and chapter ▶Bayesian Methods in Epidemiology of this handbook), an expectation is what we would or should expect to see in a given target population. This notion of expectation corresponds roughly to a prediction of what we would see if we could observe the target in question. The regression $E(Y|\boldsymbol{X}=\boldsymbol{x})$ does not represent an objective relation to be estimated but instead represents a subjective (personal) expectation about how the average of Y varies across levels of $\boldsymbol{X}$ in the target population. Like the frequentist regression estimate, however, it is something one constructs from whatever data one may find that seems informative about this variation.

When analogous models and fitting criteria are adopted in both approaches, frequentist and Bayesian methods typically yield similar interval estimates. Divergences are usually due to differences in underlying models and differences in criteria for a “good” estimate; for example, frequentists traditionally prefer unbiasedness (having an average error of zero), whereas Bayesians more often prefer closeness (e.g., having the smallest average squared error possible).

Nonetheless, even when their numeric results are the same, Bayesians and frequentists interpret these results differently. The Bayesian presents a prediction, denoted by $E(Y|\boldsymbol{X}=\boldsymbol{x})$, which he or she interprets as his or her “best bet” about the average of Y when $\boldsymbol{X}=\boldsymbol{x}$, according to some criteria for “best bet.” The frequentist presents a prediction, denoted by $\hat{E}(Y|\boldsymbol{X}=\boldsymbol{x})$ (or, more commonly, $\hat{Y}_{X=x}$), which he or she interprets as “the” best estimate of the average of Y when $\boldsymbol{X}=\boldsymbol{x}$, according to some criteria for “best estimate” (such as minimum variance among statistically unbiased estimators). Too often, the latter criteria are presumed to be universally shared, but are not really shared or even properly understood by epidemiologists; one could and would reach different conclusions using other defensible criteria (such as minimum mean squared error). For these reasons, when conducting regression analyses, it can be valuable to consider both frequentist and Bayesian interpretations of methods and results.

30.3 Basic Regression Models

In any given instance, the true regression of Y on $\boldsymbol{X}$, $E(Y|\boldsymbol{X}=\boldsymbol{x})$, is an extremely complicated function of the regressors $\boldsymbol{X}$. Thus, even if we observe this function without error, we may wish to formulate simplified pictures of reality that yield *models* for this regression. These models, while inevitably incorrect, can be very useful. A classic example is the representation of the distance from the earth to the sun, Y, as a function of day of the year T. To the nearest kilometer, this distance is a complex function of T because of the gravitational effects of the moon and of the other planets in the solar system. If we represent the orbit of the earth around the sun as a circle with the sun at the center, our regression model will predict the distance $E(Y\,|\,T=t)$ by a single number (about 150 million kilometers) that does not change with t. This model is adequate if we need only predict the distances to 2% accuracy. If we represent the orbit of the earth as an ellipse, our regression model will predict the earth-sun distance as smoothly and cyclically varying over the course of a year (within a range of about 147–153 million kilometers). Although it is not perfectly accurate, this model is adequate if we need to predict the distances to within 0.2% accuracy.

30.3.1 Model Specification and Model Fitting

Our description of the above models must be refined by distinguishing between the *form* of a model and a *fitted* model. "Circle" and "ellipse" refer to forms, that is, general classes of shapes. The circular model form corresponds to assuming a constant earth-sun distance over time; the elliptical model form allows this distance to vary over a temporal cycle. The process of deciding between these two forms is a simple example of *model specification.*

If we decide to use the circular form, we must also select a value for the radius (which is the earth-sun distance in the model). This radius specifies which circle (out of the many possible circles) to use as a representation of the earth's orbit and is an example of a model *parameter.* The process of selecting the "best" radius is an example of *model fitting,* and the circle that results is sometimes called the *fitted model* (although the latter term is sometimes used to refer to the model form instead). There are two important relations between a set of data and a model fit to those data. First, there is "distance" from the fitted model to the data; second, there is "resistance" or "stability" of the fitted model, which is the degree to which the parameter estimates change when the data themselves are changed.

Depending on our accuracy requirements, we may have on hand several simplified pictures of reality and hence several candidate models. At best, our choice might require a trade-off between simplicity and accuracy, as in the preceding example. There is an old dictum (often referred to as "Occam's razor") that one should not introduce needless complexity. According to this dictum, if we need only 2% accuracy in predicting the earth's distance from the sun, then we should not bother with the ellipse model and instead use the constant distance derived from the circle model.

There is a more subtle benefit from this advice than avoiding needless mental exertion. Suppose we are given two models, one (the more complex) containing the other (the more simple) as a special case, and some data with which to fit the two models. Then the more complex model will be able to fit the available data more closely than the simpler model, in the sense that the predictions from the more complex model will (on average) be closer to what was seen in the data than will the predictions from the simpler model. This is so in the above example because the ellipse contains the circle as a special case. Nonetheless, there is a penalty for this closeness to the data: The predictions obtained from the more complex model tend to be less stable than those obtained from the simpler model.

Consider now the use of the two different model forms to predict events outside of the data set to which the models were fit. An example would be forecasting the earth's distance from the sun; another would be predicting the incidence of AIDS 5 years in the future. Intuitively, we might expect that if one model is both closer to the data and more stable than the other, that model will give more accurate predictions. The problem is that the choice among models is rarely so clear-cut: Usually, one model will be closer to the data, while the other will be more stable, and it will be difficult to tell which will be more accurate. This is one dilemma we often face in a choice between a more complex and simpler model.

To summarize, model specification is the process of selecting a model form, while model fitting is the process of using data to estimate the parameters in a model form. There are many methods of model fitting, and the topic is so vast and technical that we will only superficially outline a few key elements. Nearly all commercial computer programs are based on one of just a few fitting methods so that nearly all users (statisticians as well as epidemiologists) are forced to base their analyses on the assumptions of these methods. We will briefly discuss specification and fitting methods below.

30.3.2 Background Example

The following epidemiological example will be used at various points to illustrate specific models. There has been controversy over whether women with no history of breast cancer but thought to be of high risk (due to family history and perhaps other factors) should be given the drug tamoxifen as a prophylactic regimen. At one point, it was suggested that tamoxifen might prevent breast cancer but also increase risk of certain other cancers.

One measure of the net impact of tamoxifen prophylaxis up to a given age is the change in risk of death by that age. Suppose the regressand Y is an indicator of death by age 70 ($Y = 1$ for dead, 0 for alive). The regressors $\boldsymbol{X}$ include:

$$\begin{aligned}
X_1 &= \text{years of tamoxifen therapy,}\\
X_2 &= \text{age (in years) at start of tamoxifen therapy,}\\
X_3 &= \text{age at menarche,}\\
X_4 &= \text{age at menopause,}\\
X_5 &= \text{parity.}
\end{aligned}$$

The target population is American women born during 1945–1950 who survive to age 50 and do not use tamoxifen before that age. If tamoxifen is not taken during follow-up, we set age at tamoxifen start (X_2) to 70 because women who start at 70 or later and women who never take tamoxifen have the same exposure history during the age interval under study.

In this example, the regression $E(Y|\boldsymbol{X}=\boldsymbol{x})$ is just the average risk, or incidence proportion, of death by age 70 among women in the target population who have $\boldsymbol{X}=\boldsymbol{x}$. Therefore, we will write $R(x)$ as a shorthand for $E(Y|\boldsymbol{X}=\boldsymbol{x})$. We will also write R for the crude (overall) average risk $E(Y)$, $R(x_1)$ for the average risk $E(Y|X_1=x_1)$ in the subpopulation defined by having $X_1=x_1$ (without regard to the other variables), and so on.

30.3.3 Vacuous Models

A model so general that implies nothing at all, but simply reexpresses the overall average risk R in a different notation, is

$$E(Y) = R = \alpha,\ 0 < \alpha < 1. \tag{30.1}$$

(This model does exclude $R=0$ or 1, but it allows R to be arbitrarily close to 0 or 1, so this exclusion is of no practical consequence.) There is only one regression parameter (or coefficient) α in this model, and it corresponds to the average risk in the target population. A model such as model 1 that has no implication (i.e., that imposes no restriction or constraint) is said to be *vacuous.*

Two models are said to be equivalent if they have identical implications for the regression. A model equivalent to model 1 is

$$E(Y) = R = \exp(\alpha),\ \alpha < 0. \tag{30.2}$$

This model has no implication. In this model, α is the natural logarithm of the overall average risk:

$$\alpha = \ln(R).$$

Another model equivalent to models 1 and 2 is

$$E(Y) = R = \text{expit}(\alpha), \tag{30.3}$$

where expit(α) is the *logistic* transform of α, defined as

$$\text{expit}(\alpha) = \frac{\exp(\alpha)}{1 + \exp(\alpha)}.$$

Again, model 3 has no implication. Now, however, the parameter α in model 3 is the logit (log odds) of the overall average risk:

$$\alpha = \ln\left(\frac{R}{1-R}\right) = \text{logit}(R).$$

For an introduction to risk measures in general see Rothman et al. (2008), Chap. 3 and chapter ►Rates, Risks, Measures of Association and Impact of this handbook.

30.3.4 Constant Models

In comparing the complexity and implications of two models A and B, we say that model A is more general, more flexible, or more complex than model B or that A contains B, if all the implications of model A are also implications of model B, but not vice versa (i.e., if B imposes some restrictions beyond those imposed by A). Other ways of stating this relation are that B is simpler, stronger, or stricter than A, B is contained or nested within A, or B is a special case of A. The following model is superficially similar to model 1, but is in fact much more strict:

$$E(Y|X_1 = x_1) = R(x_1) = \alpha \tag{30.4}$$

for all x_1. This model implies that the average risks of the subpopulations defined by years of tamoxifen use are identical. The parameter α represents the common value of these risks. This model is called a *constant* regression because it allows no variation in average risks across levels of the regressor. To see that it is a special case of model 1, note that $E(Y)$, the overall average, is just an average of all the X_1-specific averages $E(Y|X_1 = x_1)$. Hence, if all the X_1-specific averages equal α, as in model 4, then the overall average must equal α as well, as in model 1.

The following two models are equivalent to model 4:

$$R(x_1) = \exp(\alpha), \tag{30.5}$$

which can be rewritten

$$\ln[R(x_1)] = \alpha,$$

and

$$R(x_1) = \text{expit}(\alpha) = e^{\alpha}/(l + e^{\alpha}), \tag{30.6}$$

which can be rewritten

$$\text{logit}[R(x_1)] = \alpha.$$

In model 5, α is the common value of the log risks $\ln[R(x_1)]$, while in model 6, α is the common value of the logits, $\text{logit}[R(x_1)]$. Each of the equivalent models (models 4–6) is a special case of the more general models (models 1–3).

A constant regression is of course implausible in most situations. For example, age is related to most health outcomes. In the above example, we should expect the average death risk to vary across the subgroups defined by age at start (X_2). There is an infinitude of ways to model these variations. The problem of selecting a useful model from among the many choices is discussed below. For now, we only describe some of the more common choices, focusing on models for average risks (incidence proportions), incidence odds, and person-time incidence rates. The models for risks and odds can also be used to model prevalence proportions and prevalence odds.

30.3.5 Linear Risk Models

Consider the model

$$R(x_1) = \alpha + \beta_1 x_1. \tag{30.7}$$

This model allows the average risk to vary across subpopulations with different values for X_1, but only in a linear fashion. The model implies that subtracting the average risk in the subpopulation with $X_1 = x_1$ from that in the subpopulation with $X_1 = x_1 + 1$ will always yield β_1, *regardless* of what x_1 is. Under model 7,

$$R(x_1 + 1) = \alpha + \beta_1(x_1 + 1),$$

and

$$R(x_1) = \alpha + \beta_1 x_1,$$

so

$$R(x_1 + 1) - R(x_1) = \beta_1.$$

Thus, in our example, β_1 represents the difference in risk between the subpopulation defined by having $X_1 = x_1 + 1$ and that defined by having $X_1 = x_1$. The model implies that this difference does not depend on the reference level x_1 for X_1, used for the comparison.

Model 7 is an example of a *linear* risk model. It is a special case of model 1; it also contains model 4 as a special case (model 4 is the special case of model 7 in which $\beta_1 = 0$, and so average risks do not vary across levels of X_1). Linear risk models (such as model 7) are easy to understand, but have a severe technical problem that makes them difficult to fit in practice: There are combinations of α and $\beta_1$that would produce impossible values (less than 0 or greater than 1) for one or more of the risks $R(x_1)$. Several models partially or wholly address this problem by transforming the linear term $\alpha + \beta_1 x_1$ before equating it to the risk. We will study two of these models below.

30.3.6 Recentering

Under model 7,

$$R(0) = \alpha + \beta \cdot 0 = \alpha,$$

so α represents the average risk for the subpopulation with $X_1 = 0$. In the present example, 0 is a possible value for X_1 (tamoxifen), and so this interpretation of α presents no problem. Suppose, however, we modeled X_3 (age at menarche) instead of X_1:

$$R(x_3) = \alpha + \beta_3 x_3.$$

Because age at menarche cannot equal zero, α would have no meaningful interpretation in this model. In order to avoid such interpretational problems, it is a useful practice to recenter a variable for which zero is impossible (such as X_3) by subtracting some frequently observed value from it before putting it in the model. For example, age 13 is a frequently observed value for age at menarche. We can redefine X_3 to be "age at menarche minus 13 years." With this redefinition, $R(x_3) = \alpha + \beta_3 x_3$ refers to a different model, one in which $R(0) = \alpha$ represents the average risk for women who were age 13 at menarche. We will later see that such recentering is advisable when using any model and is especially important when product terms ("interactions") are used in a model.

30.3.7 Rescaling

A simple way of describing β_1 in model 7 is that it is the difference in risk per unit increase in X_1. Often the units used to measure X_1 are small relative to exposure increases of substantive interest. Suppose, for example, that X_1 was diastolic blood pressure (DBP) measured in mm Hg; β_1 would then be the risk difference per millimeter increase in DBP. A 1 mm Hg increase would, however, be of no clinical interest; instead, we would want to consider increases of at least 5 and possibly 10 or 20 mm Hg. Under model 7, the difference in risk per 10 mm Hg increase would be $10\beta_1$. If we wanted to have β_1 represent the difference in risk per 10 mm Hg, we need only redefine X_1 as DBP divided by 10; X_1 would then be DBP in cm Hg.

Division of a variable by a constant, as just described, is sometimes called *rescaling* of the variable. Such rescaling is advisable whenever it changes the measurement unit to a more meaningful value. Unfortunately, rescaling is often done in a way that makes the measurement unit *less* meaningful, by dividing the variable by its sample standard deviation (SD). The sample SD is an irregular unit unique to the study data and depends heavily on how subjects were selected into the analysis. For example, the SD of DBP might be 12.7 mm Hg in one study and 15.3 mm Hg in another study. Suppose each study divided DBP by its SD entering it in model 7. In the first study, β_1 would refer to the change in risk per 12.7 mm Hg increase in DBP, whereas in the second study, β_1 would refer to the change in risk per 15.3 mm Hg. The rescaling would thus have rendered the coefficients interpretable only in peculiar and different units so that they could not be compared directly to one another or to coefficients from other studies.

We will later see that rescaling is even more important when product terms are used in a model. We thus recommend that rescaling be done using simple and easily interpreted constants for the divisions. Methods that involve division by sample SDs (such as transformations of variables to Z-scores), however, should be avoided, whether comparing the same variables across different populations or comparing different variables in the same population (Greenland et al. 1986, 1991). In the latter case, use of quantiles (e.g., comparing risks at 90th and 10th percentiles) may be more defensible, as long as the actual values of the variable at those quantiles are given to allow comparison to other studies of the same variable.

30.3.8 Exponential Risk Models

Consider the following model:

$$R(x_1) = \exp(\alpha + \beta_1 x_1). \tag{30.8}$$

Since the exponential function (exp) is always positive, model 8 will produce positive $R(x_1)$ for any combination of α and β_1. Model 8 is sometimes called an *exponential* risk model. It is a special case of the vacuous model 2; it also contains the constant model 5 as the special case in which $\beta_1 = 0$.

To understand the implications of the exponential risk model, we can recast it in an equivalent form by taking the natural logarithm of both sides:

$$\ln[R(x_1)] = \ln[\exp(\alpha + \beta_1 x_1)] = \alpha + \beta_1 x_1. \tag{30.9}$$

Model 9 is often called a *log-linear* risk model. The exponential/log-linear model allows risk to vary across subpopulations defined by X_1, but only in an exponential fashion. To interpret the coefficients, we may compare the log risks under model 9 for the two subpopulations defined by $X_1 = x_1 + 1$ and $X_1 = x_1$:

$$\ln[R(x_1 + 1)] = \alpha + \beta_1(x_1 + 1)$$

and

$$\ln[R(x_1)] = \alpha + \beta_1 x_1,$$

so

$$\ln[R(x_1 + 1)] - \ln[R(x_1)] = \ln[R(x_1 + 1)/R(x_1)] = \beta_1.$$

Thus, under models 8 and 9, β_1 represents the log risk ratio comparing the subpopulation defined by having $X_1 = x_1 + 1$ and that defined by $X_1 = x_1$, regardless of the chosen reference level x_1. Also, ln $[R(0)] = \alpha + \beta \cdot 0 = \alpha$ if $X_1 = 0$; thus, α represents the log risk for the subpopulation with $X_1 = 0$ (and so is meaningful only if X_1 can be zero).

We can derive another (equivalent) interpretation of the parameters in the exponential risk model by noting that

$$R(x_1 + 1) = \exp[\alpha + \beta_1(x_1 + 1)]$$

and

$$R(x_1) = \exp(\alpha + \beta_1 x_1),$$

so

$$\begin{aligned} R(x_1+1)/R(x_1) &= \exp[\alpha+\beta_1(x_1+1)-(\alpha+\beta_1 x_1)] \\ &= \exp(\beta_1). \end{aligned}$$

Thus, under models 8 and 9, β_1 represents the *ratio* of risks between the subpopulations defined by $X_1 = x_1 + 1$ and $X_1 = x_1$, and this ratio does not depend on the reference level x_1 (because x_1 does not appear in the final expression for the risk ratio). Also, $R(0) = \exp(\alpha + \beta \cdot 0) = e^{\alpha}$, so e^{α} represents the average risk for the subpopulation with $X_1 = 0$.

As with linear risk models, exponential risk models have the technical problem that some combinations of α and β_1 will yield risk values greater than 1, which are impossible. This will not be a practical concern, however, if all the fitted risks and their confidence limits fall well below 1.

30.3.9 Logistic Models

Neither linear nor exponential risk models can be used to analyze case-control data if no external information is available to allow estimation of risks in the source population, whereas the following model can be used without such information:

$$\begin{aligned} R(x_1) &= \operatorname{expit}(\alpha+\beta_1 x_1) \\ &= \frac{\exp(\alpha+\beta_1 x_1)}{1+\exp(\alpha+\beta_1 x_1)}. \end{aligned} \tag{30.10}$$

This model is called a *logistic* risk model, after the logistic function (expit) in the core of its definition. Because the range of the logistic function is between 0 and 1, the model will only produce risks between 0 and 1, regardless of the values for α, β_1, and x_1. The logistic model is perhaps the most commonly used model in epidemiology, so we examine it in some detail. Model 10 is a special case of model 3, but unlike model 3, it is not vacuous because it constrains the X_1-specific risks to follow a particular (logistic) pattern. The constant model 6 is the special case of the logistic model in which $\beta_1 = 0$.

To understand the implications of the logistic model, it is helpful to recast it as a model for the odds. First, note that, under the logistic model (30.10),

$$\begin{aligned} 1-R(x_1) &= 1-\frac{\exp(\alpha+\beta_1 x_1)}{1+\exp(\alpha+\beta_1 x_1)} \\ &= \frac{1}{1+\exp(\alpha+\beta_1 x_1)}. \end{aligned}$$

Since $R(x_1)/[1-R(x_1)]$ is the odds, we divide each side of Eq. 30.10 by the last term and find that, under the logistic model, the odds of disease $O(x_1)$ when $X_1 = x_1$ is

$$O(x_1) = \frac{R(x_1)}{1 - R(x_1)} = \frac{\frac{\exp(\alpha + \beta_1 x_1)}{1 + \exp(\alpha + \beta_1 x_1)}}{\frac{1}{1 + \exp(\alpha + \beta_1 x_1)}}$$
$$= \exp(\alpha + \beta_1 x_1). \tag{30.11}$$

This equation shows that the logistic risk model is equivalent to an exponential *odds* model.

Taking logarithms of both sides of Eq. 30.11, we see that the logistic model is also equivalent to the log-linear odds model

$$\ln[O(x_1)] = \alpha + \beta_1 x_1. \tag{30.12}$$

Recall that the logit of risk is defined as the log odds:

$$\text{logit}[R(x_1)] = \ln[R(x_1)/(1 - R(x_1))] = \ln[O(x_1)].$$

Hence, from Eq. 30.12, the logistic model can be rewritten in one more equivalent form,

$$\text{logit}[R(x_1)] = \alpha + \beta_1 x_1. \tag{30.13}$$

This equivalent of the logistic model is often called the logit-linear risk model or *logit model.*

As a general caution regarding terms, note that "log-linear model" can refer to any of several different models, depending on the context: In addition to the log-linear risk model (30.9) and the log-linear *odds* model (30.12) given above, there are also log-linear *rate* models and log-linear *incidence-time* models, which will be described below.

We can derive two equivalent interpretations of the logistic model parameters. First,

$$\ln[O(x_1 + 1)] = \alpha + \beta(x_1 + 1),$$
$$\ln[O(x_1)] = \alpha + \beta_1 x_1,$$

so

$$\ln[O(x_1 + 1)] - \ln[O(x_1)] = \ln[O(x_1 + l)/O(x_1)] = \beta_1.$$

Thus, under the logistic model (30.10), β_1 represents the log odds ratio comparing the subpopulations with $X_1 = x_1 + 1$ and $X_1 = x_1$. Also, $\ln[O(0)] = \alpha + \beta_1 \cdot 0 = \alpha$; thus, α is the log odds (logit) for the subpopulation with $X_1 = 0$ (and so is meaningful only if X_1 can be zero). Equivalently, we have

$$O(x_1 + l)/O(x_1) = \exp(\beta_1),$$

and

$$O(0) = \exp(\alpha)$$

so that $\exp(\beta_1)$ is the odds ratio comparing the subpopulations with $X_1 = x_1 + 1$ and $X_1 = x_1$, and $\exp(\alpha)$ is the odds for the subpopulation with $X_1 = 0$.

Logistic models may be applied to case-control studies by reinterpreting the odds $O(x)$ as the case-control ratio in the study (Breslow and Day 1980, Chap. 6; Jewell 2004, Sect. 13.3; Greenland 2008a). For an introduction to case-control studies see Rothman et al. (2008), Chap. 8 and chapter ▸Case-Control Studies of this handbook.

30.3.10 Stratified Models and Conditional Logistic Regression

It is often the case that one wishes to use a model which is "stratified" in the sense that the intercept is allowed to vary over a series of strata. As an example, suppose we had defined 8 strata indexed by $k = 1, 2, \ldots, 8$ that represented various combinations of age and sex. A *stratified* logistic model for the dependence of risk on X_1 would then be

$$R_k(x_1) = \operatorname{expit}(\alpha_k + \beta_1 x_1).$$

Note how this differs from Eq. 30.10: Risk may now depend on the stratum k as well as on the level of X_1.

Stratified models are most commonly applied when the data have been collected with matching on the stratification variables so that each stratum k represents a particular matched set. In these applications, there may be many matched sets and thus too many intercepts to estimate using ordinary estimation techniques. A common solution is to fit a stratified logistic model using a special technique, *conditional logistic regression*, which does not estimate the intercepts but instead estimates only the regressor coefficients (β_1 above). See McCullagh and Nelder (1989), Hosmer and Lemeshow (2000), or Harrell (2001) for further details.

30.3.11 Graphical Comparison of Different Models

Suppose a particular cohort has a 1-year risk of a cardiovascular event that is 0.02 at age 50 rising to 0.32 at age 80, an absolute risk increase of 0.30, a ratio risk increase of 0.32/0.02 = 16-fold, and a ratio odds increase of $(0.32/0.68)/(0.02/0.98) = 23.06$. The average annual absolute risk increase is 0.30/30 = 0.01, but the way this increase is distributed over ages could be quite different under different models.

If the risk increase is linear in age and x is age, the linear model for the risk from age 51 to 80 would be $R(x) = \alpha_1 + \beta_1(x - 50)$. Solving $R(50) = 0.02$ and $R(80) = 0.32$, we get $\alpha_1 = 0.02$ and $\beta_1 = 0.30/30$ year $= 0.01$/year, a constant absolute increase in risk of 0.01 for each of age.

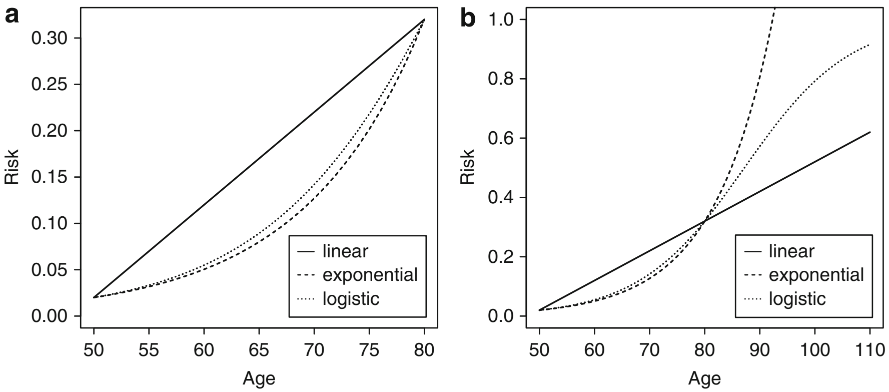


Fig. 30.1 (**a**) Risks from linear, exponential, and logistic model from age 50 to age 80 with a 1-year risk of 0.02 at age 50 and 0.32 at age 80; (**b**) risks from linear, exponential, and logistic model extrapolated to age 110 with a 1-year risk of 0.02 at age 50 and 0.32 at age 80

Now suppose the increase is exponential rather than linear. The log-linear form of the exponential model would be $\ln[R(x)] = \alpha_1 + \beta_1(x - 50)$. Solving $R(50) = 0.02$ and $R(80) = 0.32$, we now get $\alpha_1 = \ln(0.02) = -3.912$ and $\beta_1 = \ln(16)/30$ year $= 0.09242$/year, corresponding to a constant proportionate risk increase of $e^{0.09242} = 1.097$ or about 9.7% for each year of age. This corresponds to an absolute risk increase of only about 0.002 going from age 50 to 51, but of about 0.03 (15 times more) going from age 79 to 80.

Finally, suppose the increase is logistic. The logit version of the logistic model would be $\text{logit}[R(x)] = \alpha_1 + \beta_1(x - 50)$. Solving $R(50) = 0.02$ and $R(80) = 0.32$, we now get $\alpha_1 = \text{logit}(0.02) = -3.892$ and $\beta_1 = \ln(23.06)/30$ year $= 0.1046$/year, corresponding to a constant proportionate *odds* increase of $e^{0.1046} = 1.11$ or about 11% for each year of age. This corresponds to an absolute *risk* increase of only about 0.002 going from age 50 to 51, but of about 0.022 (11 times more) going from age 79 to 80.

Figure 30.1a gives plots of the risks from the above three models from age 50 to 80. The linear model produces a straight line, whereas the exponential model produces an exponential curve; these shapes will always hold when x is not transformed. The logistic curve is between the two, but is much closer in shape to the exponential for risks below 0.25, and almost the same as the exponential for risks below 10%. As shown in Fig. 30.1b, the logistic curve gradually straightens out and is close to linear for risks between 30% and 60%; above that point it begins to level off, becoming nearly flat (horizontal) as it approaches 1. In contrast, the linear and exponential curves will eventually continue on above 1 and so produce impossible values for risks (which is a problem if the actual risks could get large). For negative β_1, the curves would instead go downward from left to right.

30.3.12 Other Risk and Odds Models

In addition to those given above, several other risk models are occasionally mentioned but rarely used in epidemiology. The linear odds model is obtained by replacing the average risk by the odds in the linear risk model:

$$O(x_1) = \alpha + \beta_1 x_1. \tag{30.14}$$

Here, β_1 is the *odds* difference between subpopulations with $X_1 = x_1 + 1$ and $X_1 = x_1$, and α is the odds for the subpopulation with $X_1 = 0$. Like risk, the odds cannot be negative; unfortunately, some combinations of α and β_1 in model 14 will produce negative odds. As a result, this model (like the linear risk model) is difficult to fit and gives unsatisfactory results in many settings.

Another model replaces the logistic transform (expit) in the logistic model (30.10) by the inverse of the standard normal distribution, which also has a range between 0 and 1. The resulting model, called a*probit* model, has seen much use in bioassay. Its absence from epidemiological use may stem from the fact that (unlike the logistic) its parameters have no simple epidemiological interpretation, and the model appears to have no general advantage over the logistic in epidemiological applications.

Finally, several attempts have been made to use models that are mixtures of different basic models, especially for multiple regressions (discussed below). These mixtures have various drawbacks, including difficulties in fitting the models and interpreting the parameters (Moolgavkar and Venzon 1987). We thus do not discuss them here.

30.3.13 Rate Models

Instead of modeling average risks, we could model person-time incidence rates. If we let Y denote the *rate* observed in a study subpopulation (so that Y is the observed number of cases per unit of observed person-time), the regression $E(Y|\boldsymbol{X}=\boldsymbol{x})$ represents the average number of cases per unit of person-time in the target subpopulation defined by $\boldsymbol{X}=\boldsymbol{x}$. We will denote this expected rate or "average rate" by $I(\boldsymbol{x})$.

Most rate models are analogues of risk and odds models. For example, the model

$$I(x_1) = E(Y|X_1 = x_1) = \alpha + \beta_1 x_1 \tag{30.15}$$

is a linear *rate* model, analogous to (but different from) the linear risk and odds models (30.7 and 30.14). This rate model implies that the difference in average rates between subpopulations with $X_1 = x_1 + 1$ and $X_1 = x_1$ is β_1, regardless of x_1. Also, α is the average rate for the subpopulation with $X_1 = 0$. This model can be problematic because some combinations of α and β_1 in model 15 would produce negative rate values, which are impossible.

To prevent the latter problem, most rate modeling begins with an exponential *rate* model such as

$$I(x_1) = \exp(\alpha + \beta_1 x_1). \tag{30.16}$$

Because the exponential (exp) can never be negative, this model will not produce negative rates, regardless of α, β_1, or x_1. The model is equivalent to the log-linear *rate* model

$$\ln[I(x_1)] = \alpha + \beta_1 x_1. \tag{30.17}$$

The parameter β_1 in models 16 and 17 is the log of the rate ratio comparing the subpopulation with $X_1 = x_1 + 1$ to the subpopulation with $X_1 = x_1$, regardless of x_1; hence, $\exp(\beta_1)$ is the corresponding rate ratio $I(x_1 + 1)/I(x_1)$. Also, α is the log of the rate for the subpopulation with $X_1 = 0$; hence, $\exp(\alpha)$ is the average rate $I(0)$ when $X_1 = 0$. The exponential rate model (model 16) is analogous to, but different from, the exponential risk model (30.8) and the exponential odds model (30.11).

30.3.14 Incidence-Time Models and Hazard Models

We can also model the average time to occurrence of an event, starting from some designated zero time such as birth (in which case "time" is age), start of treatment, or some calendar date. These are called incidence-time, waiting-time, failure-time, or survival-time models (see chapter ►Survival Analysis of this handbook). Let T stand for time of the event measured from zero. One approach to incidence-time regression is to use a linear model for log incidence-time, such as

$$E[\ln(T)|X_1 = x_1] = \alpha - \beta_1 x_1. \tag{30.18}$$

Because T is always positive, $\ln(T)$ is always defined. In this model, α is the average log incidence-time in the subpopulation with $X_1 = 0$, and β_1 is the difference in average log incidence-times when comparing the subpopulation with $X_1 = x_1 + 1$ to the subpopulation with $X_1 = x_1$ (regardless of the value x_1). Model 18 is an *accelerated-life model* (Cox and Oakes 1984).

Note that the sign of β_1 in the model is reversed from its sign in earlier models. This reversal is done so that, if the outcome event at T is undesirable, then as in earlier models positive values of β_1 will correspond to harmful effects from increasing X_1, and negative values will correspond to beneficial effects. For example, under the model, if T is death time and β_1 is positive, an increase in X_1 will be associated with earlier death. Because this sign reversal is not always done, one should check what convention is followed by software used to fit model 18.

A model similar, but not identical, to model 18 is the log-linear model for expected incidence-time:

$$\ln[E(T|X_1 = x_1)] = \alpha - \beta_1 x_1. \tag{30.19}$$

Model 19 differs from model 18 because the log of an average is greater than the average of the logs (unless T does not vary). Model 19 can be rewritten

$$\begin{aligned} E(T|X_1 = x_1) &= \exp(\alpha - \beta_1 x_1) = \exp(-\beta_1 x_1)e^{\alpha} \\ &= \exp(-\beta_1 x_1)T_0, \end{aligned}$$

where $T_0 = E(T|X_1 = 0) = e^{\alpha}$. Under model 19, e^{α} is the average incidence-time in the subpopulation with $X_1 = 0$, and $e^{-\beta_1}$ is the ratio of average incidence-times in the subpopulation with $X_1 = x_1 + 1$ and the subpopulation with $X_1 = x_1$. As with model 18, the sign of β_1 is negative so that positive values of β_1 will correspond to harmful effects.

More common approaches to modeling incidence-times impose a model for the risk of the event up to each point in time or for the rate of the event at each point in time. The most famous such model is the *Cox model,* also known as the *proportional hazards model.* We can give an approximate description of this model as follows: Suppose we specify a time span Δt that is small enough so that the risk of having the event in any interval t to $t + \Delta t$ among those who survive to t without the event is very small. In its usual (exponential) form, the Cox model assumes that the rates in any such short interval will follow an exponential model like Eq. 30.16, with α, but not β_1, allowed to vary with time t.

If we write $I(t; x_1)$ for the average rate in the interval t to $t + \Delta t$ among persons who survive to t and have $X_1 = x_1$, the Cox model implies that

$$I(t; x_1) \approx \exp(\alpha_t + \beta_1 x_1). \tag{30.20}$$

Under the model, the approximation ($\approx$) improves as Δt gets smaller. Note that the intercept α_t may vary with time, but in this simple Cox model, the X_1-coefficient β_1 is assumed to remain constant. This means that, at any time t, the rate ratio comparing subpopulations with $X_1 = x_1 + 1$ and $X_1 = x_1$ will be

$$I(t; x_1 + 1)/I(t; x_1) \approx \exp[\alpha_t + \beta_1(x_1 + l)]/\exp(\alpha_t + \beta_1 x_1) = \exp(\beta_1)$$

so that β_1 is the log of the rate ratio per unit of X_1, regardless of either the reference level x_1 *or* the time t at which it is computed.

Under the Cox model (30.20), the rate at time t for the subpopulation with $X_1 = 0$ is given by $I(t; 0) = \exp(\alpha_t)$. If we denote this "baseline" rate by $\lambda_0(t)$ instead of $\exp(\alpha_t)$, we have

$$I(t; x_1) \approx \exp(\alpha_t + \beta_1 x_1) = \exp(\alpha_t + \beta_1 x_1) = \lambda_0(t)\exp(\beta_1 x_1) = \exp(\beta_1 x_1)\lambda_0(t).$$

The last expression is the standard form of the model given in most textbooks. The term "Cox model" has become fairly standard, although a special case of the model was proposed by Sheehe (1962) some 10 years before Cox (1972).

The approximate form of the Cox model (30.20) may be seen as an extension of the exponential rate model (30.16) in which the rates may vary over time. Other rate models (such as 30.15) may be used instead for the dependence of the rate on the regressor. Regardless of the particular rate model used for this dependence, the statistical theory behind the model assumes that, at each time t, the rate $I(t; x_1)$ approaches a limit $\lambda(t; x_1)$ as Δt goes to zero. This limit is usually called the *hazard* or *intensity* of the outcome at time t. A Cox model is then defined as a model in which this hazard factorizes into a term showing its dependence on the regressors and a term showing its direct dependence on time. With exponential dependence on the regressors, this factorization becomes

$$\lambda(t; x_1) = \exp(\beta_1 x_1)\lambda_0(t).$$

In epidemiological studies, these hazards are purely theoretical quantities; thus, it is important to understand the approximate forms of the model given above and what those forms imply about observable rates.

The Cox model may be extended to allow regressors to vary over time. Suppose we write $X_1(t)$ as an abbreviation for "the exposure as of time t" and $x_1(t)$ for the actual numerical value of $X_1(t)$ at time t. Then the exponential form of the *Cox model with time-dependent covariates* implies that the incidence rate at time t in the subpopulation that has exposure level $x_1(t)$ at time t is

$$I[t; x_1(t)] \approx \exp[\beta_1 x_1(t)]\lambda_0(t). \tag{30.21}$$

This model may be the most widely used model for time-dependent exposures. Usually, a time-dependent exposure $X_1(t)$ is not defined as the actual amount at time t but instead is some cumulative and lagged index of exposure up to t. For example, if time is measured in months and exposure is cumulative tamoxifen lagged 3 months, $X_1(t)$ would mean "cumulative amount of tamoxifen taken up to month $t - 3$" and $x_1(t)$ would be a value for this variable.

Serious biases can arise in use of Cox models to estimate causal effects of time-dependent exposures when time-dependent confounders are also present (Robins and Greenland 1994). Similar problems arise from the use of standard correlated-outcome ("GEE"; see chapter ▶Generalized Estimating Equations of this handbook) models for longitudinal data analysis (Robins et al. 1999). These problems can be addressed using a special class of accelerated-life models known as *structural failure-time models* (Robins et al. 1992; Robins and Greenland 1994) or by using *marginal structural models* (Robins et al. 2000). For brief descriptions of these models and their relations to conventional models, see Diggle et al. (2002) or Greenland (2008a).

30.3.15 Trend Models: Exposure Transforms

Consider again the linear risk model (30.7). If this model were correct, a plot of average risk across the subpopulations defined by X_1 (i.e., a plot of risk against X_1) would yield a line. Ordinarily, however, there is no compelling reason to think the model is correct, and we might wish to entertain other possible models for the pattern of risk traced out as exposure increases, which may involve changes in direction. We will call this pattern the *trend* in risk (in contrast to the use of "trend" to mean the general or average direction of changes in risk).

We can generate an unlimited variety of such models by *transforming* exposure, that is, by replacing X_1 in the model by some function of X_1. To illustrate, we could replace years exposed in model 7 by its logarithm, to get

$$R(x_1) = \alpha + \beta_1 \ln(x_1). \tag{30.22}$$

This is still called a linear risk model, because a plot of average risk against the new regressor $\ln(X_1)$ would yield a line. But it is a very different model from model 7 because if model 22 were correct, a plot of average risk against years exposed (X_1) would yield a *logarithmic curve* rather than a line. Such a curve starts off very steep for $X_1 < 1$, but levels off rapidly beyond $X_1 > 1$.

One technical problem can arise in using the logarithmic transform: It is not defined if X_1 is negative or zero. If the original exposure measurement can be negative or zero, it is common practice to add a number c to X_1 that is big enough to insure $X_1 + c$ is always positive. The resulting model is

$$R(x_1) = \alpha + \beta_1 1\mathrm{n}(x_1 + c). \tag{30.23}$$

The shape of the curve represented by this model (and hence results derived using the model) can be very sensitive to the value chosen for c, especially when the values of X_1 may be less than 1. Frequently, c is set equal to 1, although there is usually no compelling reason for this choice.

Among other possibilities for exposure transforms are simple power curves of the form

$$R(x_1) = \alpha + \beta_1 x_1^p, \tag{30.24}$$

where p is some number (typically 1/2 or 2) chosen in advance according to some desired property. For example, with X_1 as years exposed, use of $p = 1/2$ yields the *square-root* model

$$R(x_1) = \alpha + \beta_1 x_1^{1/2},$$

which produces a trend curve that levels off as X_1 increases above zero. In contrast, use of $p = 2$ yields the simple *quadratic* model

$$R(x_1) = \alpha + \beta_1 x_1^2,$$

which produces a trend that rises more and more steeply as X_1 increases above zero. One technical problem can arise when using the power model (30.24): It is not defined if p is fractional and X_1 can be negative. To get around this limitation, we may add some number c to X_1 that is big enough to insure $X_1 + c$ is never negative and then use $(X_1 + c)^p$ in the model; again, however, the result may be sensitive to choice of c.

The trend implications of linear and exponential models are vastly different, and hence the implications of exposure transforms are also different. Consider again the exponential risk model (30.8). If this model were correct, a plot of average risk against X_1 would yield an exponential curve, rather than a line. If β_1 is positive, this curve starts out slowly but rises more and more rapidly as X_1 increases; it eventually rises more rapidly than does any power curve (Eq. 30.24). Such rapid increase is often implausible, and we might wish to use a slower-rising curve to model risk.

One means of moderating the trend implied by an exponential model is to replace x_1 by a fixed power x_1^p with $0 < p < 1$, for example,

$$R(x_1) = \exp(\alpha + \beta_1 x_1^{1/2}).$$

Another approach is to take the logarithm of exposure. This transform produces a new model:

$$\begin{aligned} R(x_1) &= \exp[\alpha + \beta_1 1\mathrm{n}(x_1)] \\ &= \exp(\alpha)\exp[\beta_1 \ln(x_1)] \\ &= e^{\alpha}\exp[1\mathrm{n}(x_1)]^{\beta_1} = e^{\alpha} x_1^{\beta_1}. \end{aligned} \tag{30.25}$$

A graph of risk against exposure under this model produces a power curve, but now (unlike model 24), the power is the unspecified (unknown) coefficient β_1 instead of a prespecified value p, and the multiplier of the exposure power is e^{α} (which must be positive) instead of β_1. Model 25 might thus appear more appropriate than model 24 when we want the power of X_1 to appear as an unknown coefficient β_1 in the model, rather than as a prespecified value p. As earlier, however, X_1 must always be positive in order to use model 25; otherwise, one must add a constant c to it such that $X_1 + c$ is always positive. When β_1 is negative in model 25, risk declines more and more gradually across increasingly exposed subpopulations. For example, if $\beta_1 = -1$, then under Eq. 30.25 $R(x_1) = e^{\alpha}\ x_1^{-l} = e^{\alpha}/x_1$, which would imply risk declines 50% (from e^{α}/l to $e^{\alpha}/2$) when going from $X_1 = 1$ to $X_1 = 2$, but declines less than 10% (from $e^{\alpha}/10$ to $e^{\alpha}/11$) when going from $X_1 = 10$ to $X_1 = 11$.

The exposure transforms and implications just discussed carry over to the analogous models for odds and rates. For example, we can modify the logistic model (which is an exponential odds model) by substituting the odds $O(x_1)$ for the risk $R(x_1)$ in models 22–25. Similarly, we can modify the rate models by substituting the rate $I(x_1)$ for $R(x_1)$. Each model will have implications for the odds or rates analogous to those described above for the risk; because the risks, odds, and rates are functions of one another (see Rothman et al. 2008, Chap. 3), each model will have implications for other measures as well.

Any trend in the odds will appear more gradual when transformed into a risk trend. To see this, note that

$$R(x_1) = O(x_1)/[1 + O(x_1)] < O(x_1),$$

and hence,

$$O(x_1)/R(x_1) = 1 + O(x_1).$$

This ratio of odds to risk grows as the odds (and the risks) get larger. Thus, the logistic risk model, which is an exponential odds model, implies a less than exponential trend in the risk. Conversely, any trend in the risks will appear steeper when transformed into an odds trend. Thus, the exponential risk model implies a greater than exponential trend in the odds, although when risks are uniformly low (under 10% for all possible X_1 values), the risks and odds will be close and so there will be little difference between the shape of the curves produced by analogous risk and odds models.

The relation of risk and odds trends to rate trends is more complex in general, but in typical applications follows the simple rule that rate trends tend to fall between the less steep risk and more steep odds trends. For example, an exponential rate model typically implies a less than exponential risk trend but more than exponential odds trend. To see why these relations can be reasonable to expect, recall that, if incidence is measured over a span of time Δt in a closed cohort, then $R(x_1) < I(x_1)\Delta t < O(x_1)$. When the risks are uniformly low, we obtain

$R(x_1) \approx I(x_1)\Delta t \approx O(x_1)$, and so there will be little difference in the curves produced by analogous risk, rate, and odds models.

30.3.16 Interpreting Models After Transformation

One drawback of models with transformed regressors is that the interpretation of the coefficients depends on the transformation. As an example, consider the model 25, which has $\ln(x_1)$ in place of x_1. Under this model, the risk ratio for a one-unit increase in X_1 is

$$\begin{aligned} R(x_1 + 1)/R(x_1) &= e^{\alpha}(x_1 + 1)^{\beta_1}/\mathrm{e}^{\alpha}x_1^{\beta_1} \\ &= [(x_1 + 1)/x_1]^{\beta_1}, \end{aligned}$$

which will depend on the value x_1 used as the reference level: If β_1 equals 1 and x_1 is 1, the risk ratio is 2, but if β_1 equals 1 and x_1 is 2, the ratio is 1.5. Here, β_1 is the power to which x_1 is raised and so determines the shape of the trend. The interpretation of the intercept α is also altered by the transformation. Under model

25, $R(1) = e^{\alpha} 1^{\beta_1} = e^{\alpha}$; thus, α is the log risk when $X_1 = 1$, rather than when $X_1 = 0$, and so is meaningful only if 1 is a possible value for X_1.

As a contrast, consider again the model $R(x_1) = \exp(\alpha + \beta_1 x_1^{1/2})$. Use of $x_1^{1/2}$ rather than x_1 moderates the rapid increase in the slope of the exponential dose-response curve but also leads to difficulties in coefficient interpretation. Under the model, the risk ratio for a one-unit increase in X_1 is

$$\exp[\alpha + \beta_1(x_1 + 1)^{1/2}]/\exp(\alpha + \beta_1 x_1^{1/2}) = \exp\{\beta_1[(x_1 + 1)^{1/2} - x_1^{1/2}]\}.$$

Here, β_1 is the log risk ratio per unit increase in the *square root* of X_1, which is rather obscure in meaning. Interpretation may better proceed by considering the shape of the curve implied by the model, for example, by plotting $\exp(\alpha + \beta_1 x_1^{1/2})$ against possible values of X_1 for several values of β_1. (The intercept α is less important in this model because it only determines the vertical scale of the curve, rather than its shape.) Such plotting is often needed to understand and compare different transforms.

30.4 Multiple-Regression Models

Suppose now we wish to model the full multiple regression $E(Y|X = x)$. Each of the previous models for the single-regression $E(Y|X_1 = x_1)$ can be extended to handle this more general situation by using the following device: In any model for the single regression, replace $\beta_1 x_1$by

$$\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n. \tag{30.26}$$

To illustrate the idea, suppose we wish to model average risk of death by age 70 across female subpopulations defined by

$$\begin{aligned} X_1 &= \text{years of tamoxifen therapy,} \\ X_2 &= \text{age at start of tamoxifen use,} \\ X_3 &= \text{age at menarche.} \end{aligned}$$

with $\boldsymbol{X} = (X_1, X_2, X_3)$. Then the multiple linear risk model for $R(\boldsymbol{x})$ is

$$R(\boldsymbol{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

while the multiple-logistic risk model is

$$R(\boldsymbol{x}) = \text{expit}(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3).$$

If instead we wished to model the death rate, we could use the multiple linear rate model

$$I(\boldsymbol{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

or a multiple exponential rate model

$$I(\boldsymbol{x}) = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3).$$

Because formula 26 can be clumsy to write out when there are three or more regressors ($n \geq 3$), several shorthand notations are in use. Let us write $\boldsymbol{\beta}$ for the vertical list (column vector) of coefficients $\beta_1, \ldots, \beta_n$. Recall that x stands for the horizontal list (row vector) of values $x_1, \ldots, x_n$. We will let $\boldsymbol{x\beta}$ stand for $\beta_1 x_1 + \ldots + \beta_n x_n$. We can then represent the multiple linear risk model by

$$R(\boldsymbol{x}) = \alpha + \boldsymbol{x\beta} = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n, \tag{30.27}$$

the multiple-logistic model by

$$R(\boldsymbol{x}) = \text{expit}(\alpha + \boldsymbol{x\beta}), \tag{30.28}$$

the multiple exponential rate model by

$$I(\boldsymbol{x}) = \exp(\alpha + \boldsymbol{x\beta}), \tag{30.29}$$

and so on for all the models discussed earlier.

30.4.1 Relations Among Multiple-Regression Models

The multiple-regression models 27–29 are not more general than the single-regression models given earlier, nor do they contain those models as special cases. This is because they refer to entirely different subclassifications of the target population: The single-regression models refer to variations in averages across subpopulations defined by levels of just one variable; in contrast, the multiple-regression models refer to variations across the much finer subdivisions defined by the levels of several variables. For example, it is possible for $R(x_1)$ to follow the single-logistic model (30.10) without $R(\boldsymbol{x})$ following the multiple-logistic model (30.28); conversely, it is possible for $R(\boldsymbol{x})$ to follow the multiple-logistic model without $R(x_1)$ following the single-logistic model.

The preceding point is often overlooked because the single-regression models are often confused with multiple-regression models in which all regressor coefficients but one are zero. The difference is, however, analogous to the differences discussed earlier between the vacuous models 1–3 (which are so general as to imply nothing) and the constant regression models 4–6 (which are so restrictive as to be unbelievable in typical situations). To see this, consider the multiple-logistic model

$$R(\boldsymbol{x}) = \text{expit}(\alpha + \beta_1 x_1). \tag{30.30}$$

The right side of this equation is the same as in the single-logistic model (30.10), but the left side is crucially different: It is the multiple-risk regression $R(\boldsymbol{x})$, instead of the single-regression $R(x_1)$. Unlike model 10, model 30 *is* a special case of the multiple-logistic model (30.28), the one in which $\beta_2 = \beta_3 = \ldots = \beta_n = 0$. Unlike model 10, model 30 asserts that risk *does not vary* across subpopulations defined by $X_1, X_2, \ldots, X_n$ *except* to the extent that X_1 varies. This is far more strict than model 28, which allows risk to vary with $X_2, \ldots, X_n$ as well as X_1 (albeit only in a logistic fashion). It is also far more strict than model 10, which says absolutely nothing about whether or how risk varies across subpopulations defined by $X_2, \ldots, X_n$ within specific levels of X_1.

More generally, we must be careful to distinguish between models that refer to different multiple-regression functions. For example, compare the two exponential rate models:

$$I(x_1, x_2) = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2), \tag{30.31}$$

and

$$I(x_1, x_2, x_3) = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2). \tag{30.32}$$

These are different models. The first is a model for the regression of rates on X_1 and X_2 only, while the second is a model for the regression of rates on X_1, X_2, and X_3. The first model in no way refers to X_3, while the second asserts that rates do not vary across levels of X_3 if one looks within levels of X_1 and X_2. Model 32 is the special case of

$$I(x_1, x_2, x_3) = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$$

(the case in which $\beta_3 = 0$), while model 31 is not, and this special case implies model 31.

Many textbooks and software manuals fail to distinguish between models such as models 31 and 32, and instead focus only on the appearance of the right-hand side of the models. Most software fits the model that ignores other covariates (Eq. 30.31 in the above example) rather than the more restrictive model (30.32) when requested to fit a model with only X_1 and X_2 as regressors. Note that if the less restrictive model is inadequate, then the more restrictive model must also be inadequate.

Unfortunately, the converse is not true: If the less restrictive model appears adequate, it does *not* follow that the more restrictive model is also adequate. For example, it is possible for the model form $\exp(\alpha + \beta_1 x_1 + \beta_2 x_2)$ to describe adequately the double regression $I(x_1, x_2)$ (which means it describes adequately rate variation across X_1 and X_2 when X_3 is ignored) and yet at the same time describe poorly the triple regression $I(x_1, x_2, x_3)$ (which means that it describes inadequately rate variation across X_1, X_2, and X_3). That is, a model may describe poorly the rate variation across X_1, X_2, and X_3 even if it describes adequately the rate variation across X_1 and X_2 when X_3 is ignored. The decision as to whether the model is acceptable should depend on whether rate variation across X_3 is relevant to the analysis objectives. For example, if the objective is to estimate the effect of changes in X_1 on the death rate, and X_2 and X_3 are both potential confounders (as in the tamoxifen example), we would want the model to describe adequately rate variation across all three variables. But if X_3 is instead affected by the study exposure X_1 (as when X_1 is past estrogen exposure and X_3 is an indicator of current uterine bleeding), we would ordinarily not want to include X_3 in the regression model (because we would not want to adjust our exposure effect estimate for X_3).

30.4.2 Product Terms (Statistical Interactions)

Each model form described above has differing implications for measures of association derived from the models. Consider again the linear risk model with three regressors X_1, X_2, and X_3 and let x_1* and x_1 be any two values for X_1. Under the model, the risks at $X_1 = x_1$* and $X_1 = x_1$ and their difference *RD* when $X_2 = x_2$ and $X_3 = x_3$ are

$$\begin{aligned}
&R(x_1{}^*, x_2, x_3) = \alpha + \beta_1 x_1{}^* + \beta_2 x_2 + \beta_3 x_3,\\
&R(x_1, x_2, x_3) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,\\
&RD = \beta_1(x_1{}^* - x_1).
\end{aligned}$$

Thus, the model implies that the risk difference between two subpopulations with the same X_2 and X_3 levels depends only on the difference in their X_1 levels. In other words, the model implies that the risk differences for X_1 within levels of X_2 and X_3 will not vary across levels of X_2 and X_3. Such an implication may be unacceptable, in which case we can either modify the linear model or switch to another model. A simple way to modify a model is to add *product terms.* For example, suppose we want to allow the risk differences for X_1 to vary across levels of X_2. We then may add the product of X_1 and X_2 to the model as a fourth variable. The risks and their differences will then be

$$\begin{aligned}
R(x_1{}^*, x_2, x_3) &= \alpha + \beta_1 x_1{}^* + \beta_2 x_2 + \beta_3 x_3 + \gamma_{12} x_1 {*} x_2,\\
R(x_1, x_2, x_3) &= \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_{12} x_1 x_2,
\end{aligned} \tag{30.33}$$

$$RD = \beta_1(x_1{}^* - x_1) + \gamma_{12}(x_1{}^* - x_1)x_2 = (\beta_1 + \gamma_{12} x_2)(x_1{}^* - x_1). \tag{30.34}$$

Under model 33, the risk difference for $X_1 = x_1$* versus $X_1 = x_1$ is given by formula 34, which depends on X_2.

A model (e.g., 30.33) that allows variation of the risk difference for X_1 across levels of X_2 will also allow variation in the risk difference for X_2 across levels of X_1. As an example, let x_2* and x_2 be any two possible values for X_2. Under model 33, the risks at $X_2 = x_2$* and $X_2 = x_2$ and their difference *RD* when $X_1 = x_1$, $X_3 = x_3$ are

$$\begin{aligned}
&R(x_1, {x_2}^*, x_3) = \alpha + \beta_1 x_1 + \beta_2 {x_2}^* + \beta_3 x_3 + \gamma_{12} x_1 {x_2}^*,\\
&R(x_1,\ x_2,\ x_3) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_{12} x_1 x_2,\\
&RD = \beta_2({x_2}^* - x_2) + \gamma_{12} x_1 ({x_2}^* - x_2) = (\beta_2 + \gamma_{12} x_1)({x_2}^* - x_2).
\end{aligned} \tag{30.35}$$

Thus, under the model, the risk difference for $X_2 = x_2*$ versus $X_2 = x_2$ is given by formula 35, which depends on X_1. Formulas 34 and 35 illustrate how product terms modify a model in a symmetric way. The term $\gamma_{12} x_1 x_2$ allows the risk differences for X_1 to vary with X_2 and the risk differences for X_2 to vary with X_1.

A model without product terms is sometimes called a "main-effects only" model and can be viewed as the special case of a model with product terms in which all the product coefficients γ_{ij} are set to zero; it thus represents a very restrictive model. If we have three regressors in a model, we have three unique two-way regressor products ($x_1 x_2$, $x_1 x_3$, $x_2 x_3$) that we can put in the model as well. More generally, with n regressors, there are $\binom{n}{2}$ pairs and hence $\binom{n}{2}$ two-way products we can use. It is also possible to add triple products (e.g., $x_1 x_2 x_3$) or even more complex combinations to the model, but such additions are rare in practice. Notable exceptions are indices such as the body mass index (BMI) kg/m^2; however, in these situations, analysts usually neglect to examine components of the index in the model, which can introduce bias (Michels et al. 1998).

Consider next an exponential risk model with the above three variables. Under this model, the risks at $X_1 = x_1*$ and $X_1 = x_1$ and their ratio RR when $X_2 = x_2$, $X_3 = x_3$ are

$$\begin{aligned}
R({x_1}^*, x_2, x_3) &= \exp(\alpha + \beta_1 {x_1}^* + \beta_2 x_2 + \beta_3 x_3),\\
R(x_1, x_2, x_3) &= \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3),\\
RR &= \exp[\beta_1({x_1}^* - x_1)].
\end{aligned} \tag{30.36}$$

Thus, the model implies that the risk ratio comparing two subpopulations with the same X_2 and X_3 levels depends only on the difference in their X_1 levels. In other words, the model implies that the risk ratios for X_1 will be constant across levels of X_2 and X_3. If this implication is unacceptable, product terms can be inserted, as with the linear model. These terms allow the risk ratios to vary in a limited fashion across levels of other variables.

The preceding discussion of product terms can be applied to linear and exponential models in which the odds or rate replaces the risk. For example, without product terms, the logistic model implies that the odds ratios for each regressor are constant across levels of the other regressors (because the logistic model is an exponential odds model); we can add product terms to allow the odds ratios to vary. Likewise, without product terms, the exponential rate model implies that the rate ratios for each regressor are constant across levels of the other regressors; we can add product terms to allow the rate ratios to vary.

Although product terms can greatly increase the flexibility of a model, the type of variation allowed by-product terms can be very limited. For example, model 33 implies that raising X_2 by one unit (i.e., comparing subpopulations that have $X_2 = x_2 + 1$ instead of $X_2 = x_2$) will yield a risk difference for X_1 of

$$[\beta_1 + \gamma_{12}(x_2 + 1)](x_1{}^* - x_1) = (\beta_1 + \gamma_{12}x_2)(x_1{}^* - x_1) + \gamma_{12}(x_1{}^* - x_1).$$

In other words, the model implies that shifting our comparison to subpopulations that are one unit higher in X_2 will change the risk difference for X_1 in a linear fashion, by an amount γ_{12} $(x_1{}^* - x_1)$, regardless of the reference values x_1, x_2, x_3 of X_1, X_2, X_3.

30.4.3 Trends and Product Terms

Each of the above models forces or assumes a particular shape for the graph obtained when average outcome (regression) is plotted against the regressors. Consider again the tamoxifen example. Suppose we wished to plot how the risk varies across subpopulations with different number of years exposed but with the same age at start of exposure and the same age at menarche. Under the linear risk model, this would involve plotting the average risk

$$R(x_1, x_2, x_3) = \alpha + \beta x_1 + \beta_2 x_2 + \beta_3 x_3$$

against X_1, while keeping X_2 and X_3 fixed at some values x_2 and x_3. In doing so, we would obtain a line with an intercept equal to $\alpha + \beta_2 x_2 + \beta_3 x_3$ and a slope equal to β_1. Whenever we changed X_2 and X_3 and replotted $R(x)$ against X_1, the intercept would change (unless $\beta_2 = \beta_3 = 0$), but the slope would remain β_1. Because lines with the same slope are parallel, we can say that the linear risk model given above implies *parallel linear* trends in risk with increasing tamoxifen (X_1) as one moves across subpopulations of different starting age (X_2) and menarche age (X_3). This means that each change in X_2 and X_3 adds some constant (possibly negative) amount to the X_1 curve. For this reason, the linear risk model is sometimes called an *additive* risk model.

If we next plotted risks against X_2, we would get analogous results: The linear risk model given above implies parallel linear relations between average risk and X_2 as one moves across levels of X_1 and X_3. Likewise, the model implies parallel linear relations between average risk and X_3 across levels of X_1 and X_2. Thus, the linear model implies additive (parallel) relations among all the variables.

If we are unsatisfied with the linearity assumption but we wish to retain the additivity (parallel-trend) assumption, we could transform the regressors. If we are unsatisfied with the parallel-trend assumption, we can allow the trends to vary across levels of other regressors by adding product terms to the model. For example, adding the product of X_1 and X_2 to the model yields model 33, which can be rewritten

$$R(x_1, x_2, x_3) = \alpha + (\beta_1 + \gamma_{12}x_2)x_1 + \beta_2 x_2 + \beta_3 x_3.$$

From this reformulation, we see that the slope for the line obtained by plotting average risk against X_1 while keeping X_2, X_3 fixed at x_2, x_3 would be $\beta_1 + \gamma_{12}x_2$. Thus, the slope of the trend in risk across X_1 would vary across levels of X_2 (if $\gamma_{12} \neq 0$), and so the trend lines for X_1 would not be parallel. We also see that γ_{12} is the difference in the X_1-trend slopes between subpopulations with the same X_3-value but one unit apart in their X_2-value.

An entirely different approach to producing non-parallel trends begins with an exponential model. For example, under the exponential risk model (30.36), a plot of average risk against X_1 while keeping X_2 and X_3 fixed at x_2 and x_3 would produce an *exponential* curve, rather than a line. This exponential curve would have intercept $\exp(\alpha + \beta_2 x_2 + \beta_3 x_3)$. If, however, we changed the value of X_2 or X_3 and replotted risk against X_1, we would *not* obtain a parallel risk curve. Instead, the new curve would be *proportional* to the old: A change in X_2 or X_3 *multiplies* the entire X_1 curve by some amount. For this reason, the exponential model is sometimes called a *multiplicative risk* model. If we were unsatisfied with this proportionality-of-trends assumption, we could insert product terms into the model, which would allow for certain types of non-proportional trends. Proportional trends in risk appear parallel when plotted on a logarithmic vertical scale; when product terms with non-zero coefficients are present, logarithmic trends appear non-parallel.

Analogous comments and definitions apply if we substitute odds or rates for risks in the above arguments. For example, consider the multiple-logistic model in the exponential odds form:

$$O(\boldsymbol{x}) = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3).$$

A plot of the disease odds $O(\boldsymbol{x})$ against X_1 while keeping X_2 and X_3 fixed would produce an exponential curve; a plot of the log odds (logit) against X_1 while keeping X_2 and X_3 fixed would produce a line. If we changed the value of X_2 or X_3 and replotted the odds against X_1, we would obtain a new curve proportional to the old; that is, the new odds curve would equal the old multiplied by some constant amount. Thus, the logistic model is sometimes called a multiplicative-odds model. For analogous reasons, the exponential rate model is sometimes called a multiplicative-rate model. In both these models, inserting product terms into the model allows certain types of departures from proportional trends.

30.4.4 Interpreting Product-Term Models

Several important cautions should be highlighted when attempting to build models with product terms and interpret coefficients in models with product terms. First, the so-called “main-effect” coefficient β_j will be meaningless when considered alone if its regressor X_j appears in a product with another variable X_k that cannot be zero.

In the tamoxifen example, X_1 is years of exposure, which can be zero, while X_3 is age at menarche (in years), which is always above zero. Consider the model

$$\begin{aligned} R(x_1, x_2, x_3) &= \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_{13} x_1 x_3 \\ &= \alpha + \beta_1 x_1 + \beta_2 x_2 + (\beta_3 + \gamma_{13} x_1) x_3 \qquad (30.37) \\ &= \alpha + (\beta_1 + \gamma_{13} x_3) x_1 + \beta_2 x_2 + \beta_3 x_3. \end{aligned}$$

Under this model, $\beta_1 + \gamma_{13} x_3$ is the slope for the trend in risks across X_1 given $X_2 = x_2$ and $X_3 = x_3$. Thus, if X_3 was 0, this slope would be $\beta_1 + \gamma_{13} \cdot 0 = \beta_1$, and so β_1 could be interpreted as the slope for X_1 in subpopulations of a given X_2 and with $X_3 = 0$. But X_3 is age at menarche and so cannot be zero; thus, β_1 has no simple epidemiological interpretation. In contrast, because X_1 is years exposed and so can be zero, β_3 does have a simple interpretation: Under model 37, $\beta_3 + \gamma_{13} x_1$ is the slope for X_3 given $X_1 = x_1$; hence, $\beta_3 + \gamma_{13} \cdot 0 = \beta_3$ is the slope for X_3 in subpopulations with no tamoxifen exposure ($X_1 = 0$).

As mentioned earlier, if a regressor X_j cannot be zero, one can insure a simple interpretation of the intercept α by recentering the regressor, that is, by subtracting a reference value from the regressor before entering it in the model. Such recentering also helps provide a simple interpretation for the coefficients of variables that appear with X_j in product terms. In the example, we could recenter by redefining X_3 to be age at menarche -13 years. With this change, β_1 in model 37 would now be the slope for X_1 (years of tamoxifen) in subpopulations of a given X_2 (age at start of tamoxifen) in which this new X_3 was 0 (i.e., in which the age at menarche was 13).

Rescaling can also be important for interpretation of product-term coefficients. As an example, suppose X_1 is serum cholesterol in mg/dl and X_2 is diastolic blood pressure (DBP) in mm Hg and that the product of X_1 and X_2 is entered into the model without rescaling, say as $\gamma_{12} x_1 x_2$ in an exponential rate model. Then γ_{12} would represent the difference in the log of the rate ratio for a 1 mg/dl increase in cholesterol when comparing subpopulations 1 mm Hg apart in DBP. Even if this term was important, it would appear very small in magnitude because of the small units used to measure cholesterol and DBP. To avoid such deceptive appearances, we could rescale X_1 and X_2 so that their units now represented important increases in cholesterol and DBP. For example, we could redefine X_1 as cholesterol divided by 20 and X_2 as DBP divided by 10. With this rescaling, γ_{12} would represent the difference in the log of the rate ratio for a 20 mg/dl increase in cholesterol when comparing subpopulations 10 mm Hg apart in DBP.

Another caution is that, in most situations, a product term in a model should be accompanied by terms for all variables and products contained within that product. For example, if one enters $\gamma_{12} x_1 x_2$ in a model, $\beta_1 x_1$ and $\beta_2 x_2$ should also be included in that model, and if one enters $\delta_{123} x_1 x_2 x_3$ in a model, all of $\beta_1 x_1, \beta_2 x_2, \beta_3 x_3, \gamma_{12} x_1 x_2, \gamma_{13} x_1 x_3$, and $\gamma_{23} x_2 x_3$ should be included in that model. This rule, sometimes called the "hierarchy principle" (Bishop et al. 1975), is useful in avoiding models with bizarre implications. As an example, suppose X_1 is serum-lead concentration and X_2 is age minus 50 years. If $\gamma_{12} > 0$, the 1-year mortality-risk model

$$R(x_1, x_2) = \exp(\alpha + \beta_2 x_2 + \gamma_{12} x_1 x_2)$$

implies that serum-lead is positively related to risk among persons above age 50 ($X_2 > 0$), is unrelated to risk among persons of age 50 ($X_2 = 0$), and is negatively related to risk among persons below age 50 ($X_2 < 0$); if $\gamma_{12} < 0$, it implies a negative relation over 50 and a positive relation below 50. Rarely (if ever) would we have grounds for assuming such unusual relations hold. To prevent use of absurd models, many regression programs automatically enter all terms contained within a product when the user instructs the program to enter the product into the model.

Models violating the hierarchy principle often arise when one variable is not defined for all subjects or when a variable is included that is a composite of other variables (e.g., body mass index). As an example of the former, suppose in a study of breast cancer in women that X_1 is age at first birth (AFB) and X_2 is parity. Because X_1 is undefined for nulliparous women ($X_2 = 0$), one sometimes sees the breast-cancer rate modeled by a function in which age at first birth appears only in a product term with parity, such as $\exp(\alpha + \beta_2 x_2 + \gamma_1 x_1 x_2)$. The rationale for this model is that the rate will remain defined even when age at first birth (X_1) is undefined because $x_1 x_2$ will be zero when parity (X_2) is zero.

One can sometimes avoid violating the hierarchy principle if there is a reasonable way to extend variable definitions to all subjects. Thus, in the tamoxifen example, age at start of tamoxifen was extended to the untreated by setting it to age 70 (end of follow-up) for those subjects and for those subjects who started at age 70 or later. The rationale for this extension is that, within the age interval under study, untreated subjects and subjects starting tamoxifen at age 70 or later would have identical exposures.

Our final caution is that product terms are commonly labeled "interaction terms" or "statistical interactions." We avoid these labels because they may inappropriately suggest the presence of biological (mechanical) interactions between the variables in a product term. In practice, regression models are applied in many situations in which there is no effect of the regressors on the regressand (outcome). Even in causal analyses, the connections between product terms and biological interactions can be very indirect and can depend on many biological assumptions; see Rothman et al. (2008), Chap. 5 and chapter ▸Confounding and Interaction of this handbook.

30.4.5 Categorical Regressors

Consider a regressor whose possible values are discrete and few and perhaps purely nominal (i.e., with no natural ordering or quantitative meaning). An example is marital status (never married, currently married, formerly married). Such regressors may be entered into a multiple-regression model using *category indicator variables.* To use this approach, we first choose one level of the regressor as the *reference level,* against which we want to compare risks or rates. For each of the remaining levels (the *index* levels), we create a binary variable that indicates whether a person is at that level (1 if at the level, 0 if not). We then enter these indicators into the regression model.

The entire set of indicators is called the *coding* of the original regressor. To code marital status, we could take "currently married" as the reference level and define

$$\begin{aligned}X_1 &= 1 \text{ if formerly married, 0 if currently or never married,}\\ X_2 &= 1 \text{ if never married, 0 if ever married (i.e., currently or formerly married).}\end{aligned}$$

There are $2 \cdot 2 = 4$ possible numerical combinations of values for X_1 and X_2, but only three of them are logically possible. The impossible combination is $X_1 = 1$ (formerly married) and $X_2 = 1$ (never married). Note, however, that we need two indicators to distinguish the three levels of marital status because one indicator can only distinguish two levels.

In general, we need $J - 1$ indicators to code a variable with J levels. Although these indicators will have 2^{J-1} possible numerical combinations, only J of these combinations will be logically possible. For example, we will need four indicators to code a variable with five levels. These indicators will have $2^4 = 16$ numerical combinations, but only five of the 16 combinations will be logically possible.

Interpretation of the indicator coefficients depends on the model form and the chosen coding. For example, in the logistic model

$$R(x_1, x_2) = \text{expit}(\alpha + \beta_1 x_1 + \beta_2 x_2), \tag{30.38}$$

$\exp(\beta_2)$ is the odds ratio comparing $X_2 = 1$ persons (never married) to $X_2 = 0$ persons (ever married) within levels of X_1. Because one cannot have $X_2 = 1$ (never married) and $X_1 = 1$ (formerly married), the only level of X_1 within which we can compare $X_2 = 1$ to $X_2 = 0$ is the zero level (never or currently married). Thus, $\exp(\beta_2)$ is the odds ratio comparing never married ($X_2 = 1$) to currently married ($X_2 = 0$) people among those never or currently married ($X_1 = 0$). In a similar fashion, $\exp(\beta_1)$ compares those formerly married to those currently married among those ever married.

In general, the type of indicator coding described above, called *disjoint category coding,* results in coefficients that compare each index category to the reference category. With this coding, for a given person, no more than one indicator in the set can equal 1; all the indicators are zero for persons in the reference category. Another kind of coding is *nested indicator coding.* In this coding, levels of the regressor are grouped, and then codes are created to facilitate comparisons both within and across groups. For example, suppose we wish to compare those not currently married (never or formerly married) to those currently married and also compare those never married to those formerly married. We can then use the indicators

$$\begin{aligned} Z_1 &= 1 \text{ if never or formerly married (i.e., not currently married),} \\ &\quad\ 0 \text{ otherwise (currently married),} \\ Z_2 &= 1 \text{ if never married, 0 if ever married.} \end{aligned}$$

Z_2 is the same as the X_2 used above, but Z_1 is different from X_1. The combination $Z_1 = 0$ (currently married), $Z_2 = 1$ (never married) is impossible; $Z_1 = Z_2 = 1$ for people who never married. In the logistic model

$$R(z_1, z_2) = \text{expit}(\alpha + \beta_1 z_1 + \beta_2 z_2), \tag{30.39}$$

$\exp(\beta_2)$ is now the odds ratio comparing those never married ($Z_2 = 1$) to those ever married ($Z_2 = 0$) within levels of Z_1. Note that the only level of Z_1 in which this comparison can be made is $Z_1 = 1$ (never or formerly married). Similarly, $\exp(\beta_1)$ is now the odds ratio comparing those formerly married ($Z_1 = 1$) among those never married ($Z_2 = 0$).

There can be quite a large number of options for coding category indicators. The choice among these options may be dictated by which comparisons are of most interest. As long as each level of the regressor can be uniquely represented by the indicator coding, the choice of coding will not alter the assumptions represented by the model. There is, however, one technical point to consider in choosing codes. The precision of the estimated coefficient for an indicator will directly depend on the numbers of subjects at each indicator level. For example, suppose in the data there were 1,000 currently married subjects, 200 formerly married subjects, and only 10 never married subjects. Then any indicator that had "never married" as one of its levels (0 or 1) would have a much less precise coefficient estimate than other indicators. If "never married" were chosen as the reference level for a disjoint coding scheme, all the indicators would have that level as its zero level, and so all would have very imprecise coefficient estimates. To maximize precision, many analysts prefer to use disjoint coding in which the largest category (currently married in the above example) is taken as the reference level.

In choosing a coding scheme, one need not let precision concerns dominate if they get in the way of interesting comparisons. Coding schemes that distinguish among the same categories produce equivalent models. Therefore, one may fit a model repeatedly using different but equivalent coding schemes in order to easily examine all comparisons of interest. For example, one could fit model 38 to compare those never or formerly married with those currently married, then fit model 39 to compare the never with formerly married.

Although indicator coding is essential for purely nominal regressors, it can also be used to study quantitative regressors as well, especially when one expects qualitative differences between persons at different levels. Consider number of marriages as a regressor. We might suspect that people of a given age who have had one marriage tend to be qualitatively distinct from people of the same age who have had no marriage or two marriages and that people who have had several marriages are even more distinctive. We thus might want to code number of marriages in a manner that allowed qualitative distinctions among its levels. If "one marriage" was the most common level, we might take it as the reference level and use

$$X_1 = 1 \text{ if never married, 0 otherwise,}$$

$$X_2 = 1 \text{ if two marriages, 0 otherwise,}$$

$$X_3 = 1 \text{ if three or more marriages, 0 otherwise.}$$

We use one variable to represent "three or more" because there might be too few subjects with three or more marriages to produce acceptably precise coefficients for a finer division of levels. The coding just given would provide comparisons of those never married, twice married, and more-than-twice married to those once married. Other codings could be used to make other comparisons.

30.5 Trend Models with Multiple Parameters

Multiple-regression models can be extended to produce much more flexible trend models than those provided by simple transformations. The latter restrict trends to follow basic shapes, such as quadratic or logarithmic curves. The use of multiple terms for each exposure and confounder allows more detailed assessment of trends and more complete control of confounding than possible with simple transformations.

30.5.1 Categorical Trends

One way to extend trend models is to categorize the regressor and then use a category-indicator coding such as discussed above. The resulting analysis may then parallel the categorical (tabular) trend methods. To the extent allowed by the data numbers and background information, the categories should represent scientifically meaningful constructs within which risk is not expected to change dramatically. Purely mathematical categorization methods such as percentiles (quantiles) can do very poorly compared to meaningful categories. Most importantly, the choices of categories should *not* be dictated by the results produced; for example, manipulation of category boundaries to maximize the effect estimate will produce an estimate biased away from the null, while manipulation of boundaries to minimize a p-value will produce a downwardly biased p-value. Similarly, manipulation to minimize the estimate or maximize the p-value will produce a null-biased estimate or an upwardly biased p-value.

There are two common types of category codes used in trend models. *Disjoint coding* produces estimates that compare each index category (level) to the reference level. Consider coding weekly servings of fruits and vegetables with

$$X_1 = 1 \text{ for } < 15, 0 \text{ otherwise}$$
$$X_2 = 1 \text{ for } 36 - 42, 0 \text{ otherwise}$$
$$X_3 = 1 \text{ for } > 42, 0 \text{ otherwise.}$$

In the rate model

$$\ln[I(x_1, x_2, x_3)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \tag{30.40}$$

$\exp(\beta_1)$ is the rate ratio comparing the "<15" category with the "15–35" category (which is the referent), and so on, while $\exp(\alpha)$ is the rate in the "15–35" category (the category for which all the X_j are zero). When Eq. 30.40 is fit, we can plot the fitted rates on a graph as a step function. This plot provides a crude impression of the trends across (but not within) categories.

Confounders may be added to the model in order to control confounding, and these too may be coded using multiple indicators or any of the methods described below. We may plot the model-adjusted trends by fixing each confounder at a reference level and allowing the exposure level to vary.

Incremental coding (nested coding) can be useful when one wishes to compare each category against its immediate predecessor (Maclure and Greenland 1992). For "number of servings per week," we could use

$$Z_1 = 1 \text{ for } >14, 0 \text{ otherwise,}$$
$$Z_2 = 1 \text{ for } >35, 0 \text{ otherwise,}$$
$$Z_3 = 1 \text{ for } >42, 0 \text{ otherwise.}$$

Note that if $Z_2 = 1$, then $Z_1 = 1$, and if $Z_3 = 1$, then $Z_1 = Z_2 = 1$. In the model

$$\ln[I(z_1, z_2, z_3)] = \alpha + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3, \tag{30.41}$$

$\exp(\beta_1)$ is the rate ratio comparing the 15–35 category ($Z_1 = 1$ and $Z_2 = Z_3 = 0$) to the <15 category ($Z_1 = Z_2 = Z_3 = 0$). Similarly, $\exp(\beta_2)$ is the rate ratio comparing the 36–42 category ($Z_1 = Z_2 = 1$ and $Z_3 = 0$) to the 15–35 category ($Z_1 = 1$ and $Z_2 = Z_3 = 0$). Finally, $\exp(\beta_3)$ compares the >42 category ($Z_1 = Z_2 = Z_3 = 1$) to the 36–42 category ($Z_1 = Z_2 = 1$ and $Z_3 = 0$). Thus, $\exp(\beta_1)$, $\exp(\beta_2)$, and $\exp(\beta_3)$ are the incremental rate ratios across adjacent categories. Again, we may add confounders to the model and plot adjusted trends.

Another form of coding for trend analysis is called *floating absolute risk*, which avoids distortion of uncertainty assessments of trends by eliminating the need for a potentially arbitrary reference category (Easton et al. 1991; Greenland et al. 1999). In the above example, this coding would be achieved by eliminating the intercept term α from the model replacing it with an indicator $X_{1.5}$ for the (former) reference category of 15–35 servings,

$$X_{1.5} = 1 \text{ if 15–35 servings, 0 otherwise,}$$

and then using the disjoint-coding indicators for the remaining categories. The resulting model is

$$\ln[I(x_1, x_{1.5}, x_2, x_3)] = \beta_1 x_1 + \beta_{1.5} x_{1.5} + \beta_2 x_2 + \beta_3 x_3.$$

With this coding, coefficient antilogs such as $\exp(\beta_1)$ no longer represent rate ratios, but instead represent rates (if the data represent a full cohort) or expected case-control ratios (if the data represent a case-control study). The advantage is that plots of the estimated antilogs against the category scores, along with their confidence limits, will provide a more balanced picture of statistical uncertainty about the trend.

30.5.2 Regression with Category Scores

A common practice in epidemiology is to divide each covariate into categories, assign a score to each category, and enter scores into the model instead of the original variable values. Ordinal scores or codes (e.g., 1, 2, 3, 4, 5 for a series of five categories) should be avoided, as they can yield quantitatively meaningless dose-response curves and harm the power and precision of the results (Lagakos 1988; Greenland 1995b, c). Category midpoints can be much less distortive but are not defined for open-ended categories; category means or medians can be even less distortive and are defined for open-ended categories.

Unfortunately, if there are important non-linear effects within categories, no simple scoring method will yield an undistorted dose-response curve, nor will it achieve the power and precision obtainable by entering the uncategorized covariates into the model (Greenland 1995b, c). We thus recommend that when possible, categories be kept narrow and that scores be derived from category means or medians, rather than from midpoints or (worse) ordinal scores. We further recommend that one examine models with uncategorized covariates whenever effects are clearly present.

30.5.3 Power Models

Another approach to trend analysis and confounder control is to use multiple power terms for each regressor. The resulting model is usually called a *polynomial model* or *power model* (Ruppert et al. 2003). Such an approach does not require categorization but does require care in selection of terms. Traditionally, the powers used are positive integers (e.g., x_1, x_1^2, x_1^3), but fractional powers may also be used (Royston and Sauerbrei 2008). As an illustration, suppose X_1 represents the actual number of servings per week (instead of an indicator). We could model trends across this regressor by using X_1 in the model along with the following powers of X_1:

$$X_2 = X_1^{1/2} = \text{square root of } \mathrm{X}_1,$$
$$X_3 = X_1^2 \quad = \text{square of } \mathrm{X}_1.$$

The multiple-regression model

$$\ln[I(x_1, x_2, x_3)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

is now just another way of writing the power model

$$\ln[I(x_1)] = \alpha + \beta_1 x_1 + \beta_2 x_1^{1/2} + \beta_3 x_1^2. \tag{30.42}$$

We can plot fitted rates from this model using very fine spacings to produce a *smooth curve* as an estimate of rate trends across X_1. As always, we may also include confounders in the model and plot model-adjusted trends.

Power models have several advantages over categorical models. Most importantly, they make use of information about differences within categories, which is ignored by categorical models and categorical analyses (Greenland 1995a, b, c). Thus, they can provide a more complete picture of trends across exposure and more thorough control of confounders. They also provide a more smooth picture of trends. One disadvantage of power models is a potentially greater sensitivity of estimates to *outliers,* that is, persons with unusual values or unusual *combinations* of values for the regressors. As discussed below, this problem can be partially addressed by performing delta-beta analysis and by switching to other model forms such as splines.

30.5.4 Regression Splines

Often it is possible to combine the advantages of categorical and power models through the use of *spline models.* Such models can be defined in a number of equivalent ways, and we present only the simplest. In all approaches, one first categorizes the regressor, as in categorical analysis (although fewer, broader categories may be sufficient in a spline model). The boundaries between these categories are called the *knots* or *join points* of the spline. Next, one chooses the *power* (or order) of the spline, according to the flexibility one desires within the categories (higher powers allow more flexibility).

Use of category indicators corresponds to a zero-power spline, in which the trend is flat within categories but may jump suddenly at the knots; thus, category-indicator models are just special and unrealistic types of spline models. In a first-power or *linear* spline, the trend is modeled by a series of connected line segments. The trend within each category corresponds to a line segment; the slope of the trend may change only at the knots, and no sudden jump in risk (discontinuity in trend) can occur.

To illustrate how a linear spline may be represented, let X_1 again be "number of servings per week" but now define

$$X_2 = X_1 - 14 \text{ if } \mathrm{X}_1 > 14, 0 \text{ otherwise}$$
$$X_3 = X_1 - 35 \text{ if } \mathrm{X}_1 > 35, 0 \text{ otherwise}$$

Then the log-linear rate model

$$\ln[I(x_1, x_2, x_3)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{30.43}$$

will produce a log-rate trend that is a series of three line segments that are connected at the knots (category boundaries) of 14 and 35. To see this, note that when X_1 is less than 14, X_2 and X_3 are zero, so the model simplifies to a line with slope β_1:

$$\ln[I(x_1, x_2, x_3)] = \alpha + \beta_1 x_1$$

in this range. When X_1 is greater than 14 but less than 35, the model simplifies to a line with slope $\beta_1 + \beta_2$:

$$\begin{aligned}\ln[I(x_1, x_2, x_3)] &= \alpha + \beta_1 x_1 + \beta_2 x_2\\ &= \alpha + \beta_1 x_1 + \beta_2(x_1 - 14)\\ &= \alpha - 14\beta_2 + (\beta_1 + \beta_2)x_1.\end{aligned}$$

Finally, when X_1 is greater than 35, the model becomes a line with slope $\beta_1 + \beta_2 + \beta_3$:

$$\begin{aligned}\ln[I(x_1, x_2, x_3)] &= \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3\\ &= \alpha + \beta_1 x_1 + \beta_2(x_1 - 14) + \beta_3(x_1 - 35)\\ &= \alpha - 14\beta_2 - 35\beta_3 + (\beta_1 + \beta_2 + \beta_3)x_1.\end{aligned}$$

Thus, β_1 is the slope of the spline in the first category, β_2 is the change in slope in going from the first to second category, and β_3 is the change in slope in going from the second to third category.

The trend produced by a linear spline is generally more realistic than a categorical trend, but can suddenly change its slope at the knots. To smooth out such sudden changes, we may increase the order of the spline. Increasing the power to 2 produces a second-power or *quadratic* spline, which comprises a series of parabolic curve segments smoothly joined together at the knots. To illustrate how such a trend may be represented, let X_1, X_2, and X_3 be as just defined. Then the model

$$\ln[I(x_1, x_2, x_3)] = \alpha + \beta_1 x_1 + \gamma_1 x_1^2 + \gamma_2 x_2^2 + \gamma_3 x_3^2 \tag{30.44}$$

will produce a log-rate trend that is a series of three parabolic segments smoothly connected at the knots of 14 and 35. The coefficient γ_1 corresponds to the curvature of the trend in the first category, while γ_2 and γ_3 correspond to the changes in curvature when going from the first to second and second to third category. A still smoother curve could be fit by using a third-power or *cubic* spline, but for epidemiological purposes the quadratic spline is often smooth and flexible enough.

Although for simplicity it was not done in the above examples, as mentioned earlier, it is best to recenter quantitative variables before model fitting. For splines, this will require recentering the knots as well. Thus, suppose we begin analysis by subtracting 20 from number of servings X_1. The knots to create X_2 and X_3 from this recentered X_1 (which is number of servings −20) must then also be recentered to become $14 - 20 = -6$ and $35 - 20 = 15$. With this recentering $X_1 = X_2 = X_3 = 0$, and so the intercept α will be the log-rate at 20 servings.

One disadvantage of quadratic and cubic splines is that the curves in the end categories (tails) may become very unstable, especially if the category is open-ended. This instability may be reduced by *restricting* one or both of the end categories to be a line segment rather than a curve. To restrict the lower category to be linear in a quadratic spline, we need only drop the *first* quadratic term $\gamma_1 x_1^2$ from the model; to restrict the upper category, we must subtract the *last* quadratic term from all the quadratic terms and drop the last term out of the model. To illustrate an upper category restriction, suppose we wish to restrict the above quadratic spline model for log-rates (Eq. 30.44) so that it is linear in the upper category only. Define

$$\begin{aligned} Z_1 &= X_1 = \text{number of servings per week}, \\ Z_2 &= X_1^2 - X_3^2, \\ Z_3 &= X_2^2 - X_3^2. \end{aligned}$$

Then the model

$$\ln[I(z_1, z_2, z_3)] = \alpha + \beta_1 z_l + \beta_2 z_2 + \beta_3 z_3 \tag{30.45}$$

will produce a log-rate trend that comprises smoothly connected parabolic segments in the first two categories ("<14" and "15–35"), and a line segment in the last category (">35") that is smoothly connected to the parabolic segment in the second category. (If we also wanted to force the log-rate curve in the first category to follow a line, we would drop Z_2 from the model.)

To plot or tabulate a spline curve from a given spline model, we select a set of X_1 values spaced across the range of interest, compute the set of spline terms for each X_1 value, combine these terms with the coefficients in the model to get the model-predicted outcomes, and plot these predictions. To illustrate, suppose X_1 is servings per week and we wish to plot model 45 with $\alpha = -6.00$, $\beta_1 = -0.010$, $\beta_2 = -0.001$, and $\beta_3 = 0.001$ over the range 0–50 servings per week in 5-serving increments. We then compute Z_1, Z_2, Z_3 at $0, 5, 10, \ldots, 50$ servings per week and compute the predicted rate

$$\exp(-6.00 - 0.010z_1 - 0.001z_2 + 0.001z_3)$$

at each set of Z_1, Z_2, Z_3 values and plot these predictions against the corresponding X_1 values 0, 5, 10, ..., 50. For example, at $X_1 = 40$, we get $Z_1 = 40$, $Z_2 = 40^2$–$(40–35)^2 = 1575$, and $Z_3 = (40–14)^2$–$(40–35)^2 = 651$ for a predicted rate of

$$\exp[-6.00 - 0.010(40) - 0.001(1575) + 0.001(651)] = 2/1000 \text{ year.}$$

As with other trend models, we may obtain model-adjusted trends by adding confounder terms to our spline models. The confounder terms may be splines or any other form we prefer; spline plotting will be simplified, however, if the confounders are centered before they are entered into the analysis, for then the above plotting method may be used without modification.

For discussion of splines and their extensions and relations to more general regression techniques, see, for example, Hastie et al. (2009), Harrell (2001), or Ruppert et al. (2003).

30.5.5 Exclusion of Individuals from Trend Fitting

In many settings, there may be concern that individuals at a particular exposure level are qualitatively different from other individuals with respect to unmeasured or poorly measure risk factors. If such differences exist, inclusion of these individuals in the analysis will distort a fitted trend curve away from the true causal dose-response form. For example, in studies of alcohol and health outcomes, there may be concern that non-drinkers of alcohol include many individuals who do not drink due to poor health, thus creating elevated risk among non-drinkers and the confounded appearance of a protective effect of alcohol.

A simple way to address such concerns is to exclude such individuals from trend estimation. This exclusion will however reduce the number of individuals available for measured-confounder control (note that the effects of measured confounders need not be estimated in an unbiased fashion in order to control confounding; they may reflect unmeasured confounder effects as well). An alternative is to then include the individuals along with an indicator variable that separates them from the trend curve. In the alcohol example, this would involve adding an indicator for "non-drinker" to the model. The resulting fitted trend for alcohol would exclude non-drinkers, but the latter would still contribute to estimating, for example, age and sex effects. For further discussion of this strategy, see Greenland and Poole (1995).

30.5.6 Models for Trend Variation

We may allow trends to vary across regressor levels by entering products among regressor terms. For example, suppose X_1, X_2, X_3 are power terms for fruit and vegetable intake, while W_1, W_2, W_3, and W_4 are spline terms for age. To allow the fruit-vegetable trend in log-rates to vary with age, we could enter into the model all $3 \cdot 4 = 12$ products of the X_j and W_k, along with the X_j and W_k. If in addition there was an indicator $Z_1 = 1$ for female, 0 for males, the resulting model would be

$$\begin{aligned}
&\ln[R(x_1, x_2, x_3, w_1, w_2, w_3, w_4, z_1)]\\
&=\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 w_1 + \beta_5 w_2 + \beta_6 w_3 + \beta_7 w_4 + \beta_8 z_1\\
&\quad + \gamma_{11} x_1 w_1 + \gamma_{12} x_1 w_2 + \ldots + \gamma_{33} x_3 w_3 + \gamma_{34} x_3 w_4.
\end{aligned}$$

The same model form may be used if X_1, X_2, and X_3 and W_1, W_2, W_3, and W_4 represent category indicators or other terms for fruit-vegetable intake and age.

Models with products among multiple trend terms can be difficult to fit and may yield quite unstable results unless large numbers of cases are observed. Given enough data, however, such models can provide more realistic pictures of dose-response relations than can simpler models. Results from such models may be easily interpreted by plotting or tabulating the fitted trends for the key exposures of interest at various levels of the “modifying” regressors. In the above example, this process would involve plotting the model-fitted rates against fruit and vegetable intake for each of several ages (e.g., for ages evenly spaced within the range of case ages).

30.6 Extensions of Logistic Models

Outcomes that are polytomous or continuous are often analyzed by reducing them to just two categories and applying a logistic model. For example, CD4 counts might be reduced to the dichotomy $\leq$200, $>$200; cancer outcomes might be reduced to cancer and no cancer. Alternatively, multiple categories may be created with one designated as a referent and the other categories compared one at a time to the referent using separate logistic models for each comparison. While not necessarily invalid, these approaches disregard the information contained in differences within categories, in differences between non-reference categories, and in ordering among the categories. As a result, models specifically designed for polytomous or continuous outcomes can yield more precision and power than simple dichotomous-outcome analyses.

This section briefly describes several extensions of the multiple-logistic model (30.28) to polytomous and ordinal outcomes. Analogous extensions of other models are possible.

30.6.1 Polytomous Logistic Models

Suppose an outcome variable Y has $I+1$ mutually exclusive outcome categories or levels $y_0, \ldots, y_I$, where category y_0 is considered the reference category. For example, in a case-control study of relations of exposures to types of cancer, Y is a disease outcome variable, with $y_0 =$ all control as the reference category, and I other categories $y_1, \ldots, y_I$, which correspond to the cancer outcomes (leukemia, lymphoma, lung cancer, etc.). Let $R_i(\boldsymbol{x})$ denote the average risk of falling in outcome category $Y_i (i = 1, \ldots, I)$ given that the regressors $\boldsymbol{X}$ equal $\boldsymbol{x}$; that is, let

$$R_i(\boldsymbol{x}) = \Pr(Y = y_i | \boldsymbol{X} = \boldsymbol{x}).$$

The *polytomous logistic model* for this risk is then

$$R_i(\boldsymbol{x}) = \frac{\exp(\alpha_i + x\beta_i)}{1 + \sum\limits_{j=1}^{I} \exp(\alpha_j + x\beta_j)}. \tag{30.46}$$

This is a model for the risk of falling in cancer category y_i. When Y has only two levels, I equals 1, and so formula 46 simplifies to the binary multiple-logistic model (30.28).

Model 46 represents I separate logistic risk equations, one for each non-reference outcome level $y_1, \ldots, y_I$. Each equation has its own intercept α_i and vector of coefficients $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{in})$ so that there is a distinct coefficient β_{ik} corresponding to every combination of a regressor X_k and non-reference outcome level $y_i (i = 1, \ldots, I)$. Thus, with n regressors in $\boldsymbol{X}$, the polytomous logistic model involves I intercepts and $I \cdot n$ regressor coefficients. For example, with seven non-reference outcome levels and three regressors, the model would involve seven intercepts and $7 \cdot 3 = 21$ regressor coefficients, for a total of 28 model parameters.

The polytomous logistic model can be written more simply as a model for the odds. To see this, note that the risk of falling in the reference category must equal one minus the sum of the risks of falling in the non-reference categories:

$$\begin{aligned} R_0(\boldsymbol{x}) &= \Pr(Y = y_0 | \boldsymbol{X} = \boldsymbol{x}) \\ &= 1 - \sum_{i=1}^{I} \exp(\alpha_i + \boldsymbol{x\beta}_i) / [1 + \sum_{j=1}^{I} \exp(\alpha_j + \boldsymbol{x\beta}_j)] \\ &= 1/[1 + \sum_{j=1}^{I} \exp(\alpha_j + \boldsymbol{x\beta}_j)]. \end{aligned} \tag{30.47}$$

Dividing Eq. 30.47 into Eq. 30.46, we get a model for $O_i(\boldsymbol{x}) = R_i(\boldsymbol{x})/R_0(\boldsymbol{x})$ = the odds of falling in outcome category y_i versus category y_0:

$$\begin{aligned} O_i(\boldsymbol{x}) &= \frac{\exp(\alpha_i + \boldsymbol{x\beta}_i)/[1 + \sum_j \exp(\alpha_j + \boldsymbol{x\beta}_j)]}{1/[1 + \sum_j \exp(\alpha_j + \boldsymbol{x\beta}_j)} \\ &= \exp(\alpha_i + \boldsymbol{x\beta}_i). \end{aligned} \tag{30.48}$$

This form of the model shows that the polytomous model is equivalent to assuming I binary logistic models comparing each of the I non-reference categories to the reference category y_0 (see Eq. 30.11). The advantage of combining the models into one formula comes in the fitting stage: Fitting Eq. 30.46 produces more precise estimates than fitting I equations separately.

Equation 30.48 also provides a familiar interpretation for the coefficients. Suppose $\boldsymbol{x}_1$ and $\boldsymbol{x}_0$ are two different vectors of values for the regressors $\boldsymbol{X}$. Then the ratio of the odds of falling in category y_i versus y_0 when $\boldsymbol{X} = \boldsymbol{x}_1$ and $\boldsymbol{X} = \boldsymbol{x}_0$ is

$$\frac{O_i(\boldsymbol{x}_1)}{O_i(\boldsymbol{x}_0)} = \frac{\exp(\alpha_i + \boldsymbol{x}_1\boldsymbol{\beta}_i)}{\exp(\alpha_i + \boldsymbol{x}_0\boldsymbol{\beta}_i)} = \exp[(\boldsymbol{x}_1 - \boldsymbol{x}_0)\boldsymbol{\beta}_i].$$

From this equation, we see that the antilog $\exp(\beta_{ik})$ of a coefficient β_{ik} corresponds to the proportionate change in the odds of outcome i when the regressor X_k increases by one unit.

The polytomous logistic model is most useful when the levels of Y have no meaningful order, as with the cancer types. For further reading about the model, see McCullagh and Nelder (1989) and Hosmer and Lemeshow (2000).

30.6.2 Ordinal Logistic Models

Suppose that the levels $y_0, \ldots, y_I$ of Y follow a natural order. Order arises, for example, when Y is a clinical scale, such as y_0 = normal, y_1 = dysplasia, y_2 = neoplasia, rather than just a cancer indicator; Y is a count, such as number of malformations found in an individual; or the Y levels represent categories of a physical quantity, such as CD4 count (e.g., >500, 200–500, <200). There are at least four different ways to extend the logistic model to such outcomes.

Recall that the logistic model is equivalent to an exponential odds model. The first extension uses an exponential model to represent the odds of falling in outcome category y_i versus falling in category y_{i-1} (the next lowest category):

$$\frac{R_i(\boldsymbol{x})}{R_{i-1}(\boldsymbol{x})} = \frac{\Pr(Y = y_i|\boldsymbol{X} = \boldsymbol{x})}{\Pr(Y = y_{i-1}|\boldsymbol{X} = \boldsymbol{x})} = \exp({\alpha_i}^* + \boldsymbol{x\beta}^*) \tag{30.49}$$

for $i = 1, \ldots, I$. This may be called the *adjacent-category logistic model,* because taking logarithms of both sides yields the equivalent *adjacent-category logit* model

(Agresti 2002). It is a special case of the polytomous logistic model: From Eq. 30.48, the polytomous logistic model implies that

$$\frac{R_i(x)}{R_{i-1}(x)} = \frac{R_i(x)/R_0(x)}{R_{i-1}(x)/R_0(x)} = \frac{\exp(\alpha_i + \boldsymbol{x\beta}_i)}{\exp(\alpha_{i-1} + \boldsymbol{x\beta}_{i-1})}$$
$$= \exp[(\alpha_i - \alpha_{i-1}) + \boldsymbol{x}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i-1})].$$

The adjacent-category logistic model sets $\alpha_i{}^* = \alpha_i – \alpha_{i-1}$ and forces the I coefficient differences $\boldsymbol{\beta}_i – \boldsymbol{\beta}_{i-1} (i = 1,\ldots, I)$ to equal a common value $\boldsymbol{\beta}^*$. If there is a natural distance d_i between adjacent outcome categories y_i and y_{i-1} (such as the difference between the category means), the model can be modified to use these distances as follows:

$$R_i(\boldsymbol{x})/R_{i-1}(\boldsymbol{x}) = \exp(\alpha_i{}^* + \boldsymbol{x\beta}^* d_i) \tag{30.50}$$

for $i = 1, \ldots, I$. This model allows the coefficient differences $\boldsymbol{\beta}_i – \boldsymbol{\beta}_{i-1}$ to vary with the distances d_i between categories.

The second extension uses an exponential model to represent the odds of falling *above category* y_i versus*falling in or below* category y_i:

$$\frac{\Pr(Y > y_i|\boldsymbol{X} = \boldsymbol{x})}{\Pr(Y \le y_i|\boldsymbol{X} = \boldsymbol{x})} = \exp(\alpha_i{}^* + \boldsymbol{x\beta}^*), \tag{30.51}$$

where $i = 0, \ldots, I$. This is called the *cumulative-odds* or *proportional-odds* model. It can be derived by assuming that Y was obtained by categorizing a special type of continuous variable.

The third extension uses an exponential model to represent the odds of falling *above outcome* category y_i versus *in* category y_i:

$$\frac{\Pr(Y > y_i|\boldsymbol{X} = \boldsymbol{x})}{\Pr(Y = y_i|\boldsymbol{X} = \boldsymbol{x})} = \exp(\alpha_i^* + \boldsymbol{x\beta}^*), \tag{30.52}$$

where $i = 0, \ldots, I$. This is called the *continuation-ratio model.* When Y represents the time of disease incidence, Eq. 30.52 is the discrete-time analogue of the Cox model (30.20) but with coefficient signs reversed.

The fourth extension uses an exponential model to represent the odds of falling *in* category y_i versus falling *below* y_i:

$$\frac{\Pr(Y = y_i|\boldsymbol{X} = \boldsymbol{x})}{\Pr(Y < y_i|\boldsymbol{X} = \boldsymbol{x})} = \exp(\alpha_i^* + \boldsymbol{x\beta}^*), \tag{30.53}$$

where $i = 1, \ldots, I$. This model may be called the *reverse continuation-ratio model.* It can be derived by reversing the order of the Y levels in model (30.52), but in any given application, it is not equivalent to model (30.52) (Greenland 1994).

How does one choose from the variety of ordinal models? Certain guidelines may be of use, although none is compelling. First, the adjacent-category and cumulative-odds models are *reversible* in that only the signs of the coefficients change if the order of the Y levels is reversed. In contrast, the two continuation-ratio models are not reversible. This observation suggests that the continuation-ratio models may be more appropriate for modeling irreversible disease stages (e.g., osteoarthritic severity), whereas the adjacent-category and cumulative-odds models may be more appropriate for potentially reversible outcomes (e.g., blood pressure, cell counts) (Greenland 1994). Second, because the coefficients of adjacent-category models contrast pairs of categories, the model appears best suited for discrete outcomes with few levels (e.g., cell types along a normal-dysplastic-neoplastic scale). Third, because the cumulative-odds model can be derived from categorizing certain special types of continuous outcomes, it is often considered most appropriate when the outcome under study is derived by categorizing a single underlying continuum (e.g., blood pressure) (McCullagh and Nelder 1989). For a more detailed comparative discussion of ordinal logistic models and guidelines for their use, see Greenland (1994) and Ananth and Kleinbaum (1997).

All the above ordinal models simplify to the ordinary logistic model when there are only two outcome categories ($I = 2$). One advantage of the continuation-ratio models over their competitors is of special importance: Estimation of the coefficients $\boldsymbol{\beta}*$ in those models can be carried out if the levels of Y are numerous and sparse; Y may even be continuous. Thus, one can apply the continuation-ratio models without any categorization of Y. This advantage can be important because results from all the above models (including the cumulative-odds model) may be affected by the choice of the Y categories (Greenland 1994; Strömberg 1996). The only caution is that conditional (as opposed to unconditional) maximum likelihood must be used to fit the continuation-ratio model if the observed outcomes are sparsely scattered across the levels of Y (as would be inevitable if Y were continuous). See Greenland (1994) for further details and Cole and Ananth (2001) for further extensions of the model.

30.7 Generalized Linear Models

Consider again the general form of the exponential risk and rate models, $R(\boldsymbol{x}) = \exp(\alpha + \boldsymbol{x}\boldsymbol{\beta})$ and $I(\boldsymbol{x}) = \exp(\alpha + \boldsymbol{x}\boldsymbol{\beta})$ and the logistic risk model $R(\boldsymbol{x}) = \text{expit}$ $(\alpha + \boldsymbol{x}\boldsymbol{\beta})$. There is no reason why we cannot replace the "exp" in the exponential models or the "expit" in the logistic model by some other reasonable function. In fact, each of these models is of the general form

$$E(Y|\boldsymbol{x}) = f(\alpha + \mathbf{x}\boldsymbol{\beta}), \tag{30.54}$$

where f is some function that is smooth and strictly increasing (i.e., as $\alpha + \boldsymbol{x}\boldsymbol{\beta}$ gets larger, $f(\alpha + \boldsymbol{x}\boldsymbol{\beta})$ gets larger, but never jumps or bends suddenly).

For any such function f, there is always an inverse function g that "undoes" f, in the sense that $g[f(u)] = u$ whenever $f(u)$ is defined. Hence, a general form equivalent to Eq. 30.54 is

$$g[E(Y|\boldsymbol{x})] = \alpha + \boldsymbol{x\beta}. \tag{30.55}$$

A model of the form Eq. 30.55 is called a *generalized linear model.* The function g is called the *link function* for the model; thus, the link function is ln for the log-linear model and logit for the logit-linear model. The term $\alpha + \boldsymbol{x\beta}$ in it is called the *linear predictor* for the model and is often abbreviated η (Greek letter eta); that is, $\eta = \alpha + \boldsymbol{x\beta}$ by definition.

Subject to some technicalities concerning how to define the expected values, all the models we have discussed can be viewed as generalized linear models or transforms of them. Ordinary linear models (such as the linear risk model) are the simplest examples, in which *f* and *g* are both the identity function $f(u) = g(u) = u$, so that

$$E(Y|x) = \alpha + \boldsymbol{x\beta}.$$

The inverse of the exponential function exp is the natural-log function ln(u). Hence, the generalized linear forms of the exponential risk and rate models are the log-linear risk and rate models

$$\ln[R(\boldsymbol{x})] = \alpha + \boldsymbol{x\beta} \quad \text{and} \quad \ln[I(\boldsymbol{x})] = \alpha + \boldsymbol{x\beta};$$

that is, the exponential risk and rate models correspond to a natural-log link function because ln[exp(u)] = u. Similarly, the inverse of expit, the logistic function, is the logit function logit (u). Hence, the generalized linear form of the logistic risk model is the logit-linear risk model

$$\text{logit}[R(\boldsymbol{x})] = \alpha + \boldsymbol{x\beta};$$

that is, the logistic model corresponds to the logit link function because logit[expit(u)] = u.

The choices for f and g are virtually unlimited. In epidemiology, however, only the logit link $g(u) = \text{logit}(u)$ is in common use for risks, and only the log link $g(u) = \ln(u)$ is in common use for rates. In practice, these link functions are almost always the default and are sometimes the only options in commercial software for risk and rate modeling. Some packages, however, allow easy selection of linear risk, rate, and odds models, which use the identity link. Some software (e.g., GLIM) allows the user to define their own link function.

The choice of link function can have a profound impact on the shape of the trend or dose-response surface allowed by the model, especially if exposure is represented by only one or two terms. For example, if exposure is represented by a single term $\beta_1 x_1$ in a risk model, use of the identity link results in a linear risk model and a linear trend for risk, use of the log link results in an exponential (log-linear) risk model and an exponential trend for risk, and use of a logit link results in a logistic model and an exponential trend for the odds. Generalized linear models encompass a broader range than the linear, log-linear, and logistic forms, however. One example is the complementary log-log risk model,

$$R(x) = 1 - \exp[-\exp(\alpha + \boldsymbol{x\beta})],$$

which translates to the generalized linear form

$$\ln[-\ln(l - R(\boldsymbol{x}))] = \alpha + \boldsymbol{x\beta}.$$

This model corresponds to the link function ln[−ln(l−u)] and arises naturally in certain biology experiments. For further reading on this and other generalized linear models, see McCullagh and Nelder (1989).

30.8 Model Searching

How do we find a model or set of models acceptable for our purposes? There are far too many model forms to allow us to examine most or even much of the total realm of possibilities. There are several systematic, mechanical, and traditional algorithms for finding models (such as stepwise and best-subset regression) that lack logical or statistical justification and that perform poorly in theoretical, simulation, and case studies; see Sclove et al. (1972), Bancroft and Han (1977), Draper et al. (1979), Freedman (1983), Flack and Chang (1987), Freedman et al. (1988), Hurvich and Tsai (1990), Faraway (1992), Weiss (1995), Greenland (1993, 2000a, 2001), Harrell (2001), Viallefont et al. (2001), and Steyerberg (2009, Chap. 11).

One serious problem is that the p-values and standard errors obtained when variables are selected using significance-testing criteria (such as "F-to-enter" and "F-to-remove") will be downwardly biased, usually to a large degree. In particular, standard errors obtained from the selected model underestimate the standard deviations of the point estimates obtained by applying the algorithms across different random samples. As a result, the algorithms will tend to yield p-values that are too small and confidence intervals that are too narrow (and hence fail to cover the true coefficient values with the stated frequency). Unfortunately, significance-testing criteria are the basis for most variable-selection procedures in standard packaged software.

Other criteria for selecting variables, such as "change-in-point-estimate" criteria, do not necessarily perform better than significance testing (Maldonado and Greenland 1993a). Viable alternatives to significance testing in model selection have emerged only gradually with recent advances in computing and with deeper insights into the problem of model selection. We first outline the traditional approaches after reinforcing one of the most essential and neglected starting points for good modeling: laying out existing information in a manner that can help the search avoid models in conflict with established facts. A powerful alternative to model selection is provided by hierarchical regression, also known as multilevel, mixed-model, penalized, ridge, shrinkage, or random-coefficient regression (Greenland 2000a, b; Harrell 2001; Ruppert et al. 2003).

30.8.1 Role of Prior Information

The dependence of regression results on the chosen model can be either an advantage or a drawback. The advantage comes from the fact that use of a model structure capable of reasonably approximating reality can elevate the accuracy of the estimates over those from the corresponding tabular analysis. The drawback comes from the fact that use of a model incapable of even approximating reality can decrease estimation accuracy below that of tabular analysis.

This duality underscores the desirability of using flexible (and possibly complex) models. One should take care to avoid models that are entirely unsupported by background knowledge. For example, in a cohort study of lung cancer, it is reasonable to restrict rates to increase with age because there is enormous background literature documenting that this trend is found in all human populations. In contrast, one would want to avoid restricting cardiovascular disease (CVD) rates to strictly increase with alcohol consumption because there are considerable data to suggest the alcohol-CVD relation is not strictly increasing (Maclure 1993).

Prior knowledge about most epidemiological relations is usually too limited to provide much guidance in model selection. A natural response might be to use models as flexible as possible (a flexible model can reproduce a wide variety of curves and surfaces). Unfortunately, flexible models have limitations. The more flexible the model, the larger the sample needed for the usual estimation methods (such as maximum likelihood) to provide approximately unbiased coefficient estimates. Also, after a certain point, increasing flexibility may increase variability of estimates so much that the accuracy of the estimates is decreased relative to estimates from simpler models, despite the greater faithfulness of the flexible model to reality. As a result, it is usual practice to employ models that are severely restrictive in arbitrary ways, such as models without product terms or models that set "non-significant" coefficients to zero (Robins and Greenland 1986). Hierarchical methods can help alleviate some of these problems by allowing one to fit larger models than one can with ordinary methods (Greenland 2000b, 2008a, b).

Fortunately, estimates obtained from the most common epidemiological regression models, exponential (log-linear) and logistic models, retain some interpretability even when the underlying (true) regression function is not particularly close to those forms (Maldonado and Greenland 1993b; Greenland and Maldonado 1994). For example, under reasonably common conditions, rate-ratio or risk-ratio estimates obtained from those models can be interpreted as approximate estimates of standardized rate or risk ratios, using the total source population as the standard (Greenland and Maldonado 1994). To ensure such interpretations are reasonable, the model used should at least be able to replicate qualitative features of the underlying regression function. For example, if the underlying regression may have a reversal in the slope of the exposure-response curve, we should want to use a model capable of exhibiting such reversal (even if it cannot replicate the exact shape of the true curve).

A major problem for epidemiology is that key variables may be unmeasured or poorly measured. No conventional method can account for these problems. Unmeasured variables may be modeled using prior information on their relation to measured variables, but the results will be entirely dependent on that information (Leamer 1978; Greenland 2003, 2005b, 2009a, b; Gustafson 2005). Occasionally, measurement-error information may be in the form of data that can be used in correction techniques (Gustafson 2003; Carroll et al. 2006; Greenland and Lash 2008; cf. chapter ▸Measurement Error of this handbook); otherwise, sensitivity analyses or special modeling methods will be needed (Greenland and Lash 2008; Greenland 2005b, 2009a, b; Lash et al. 2009; cf. chapter ▸Sensitivity Analysis and Bias Analysis of this handbook).

30.8.2 Selection Strategies

Even with ample prior information, there will always be an overwhelming number of model choices, and so model search strategies will be needed. Many strategies have been proposed, although none has been fully justified.

Some strategies begin by specifying a minimal model form that is among the most simple credible forms. Here "credible" means "compatible with available information." Thus, we start with a model of minimal computational or conceptual complexity that does not conflict with background information. There may be many such models; in order to help insure that our analysis is credible to the intended audience, however, the starting model form should be one that most researchers would view as a reasonable possibility.

To specify a simple yet credible model form, one needs some knowledge of the background scientific literature on the relations under study. This knowledge would include information about relations of potential confounders to the study exposures and study diseases as well as relations of study exposures to the study diseases. Thus, specification of a simple yet credible model can demand much more initial effort than is routinely used in model specification.

Once we have specified our minimal starting model, we can add complexities or terms that seem necessary in light of the data. Common criteria for adding terms are discussed below in Sect. 30.10.3 "Tests of Fit." Such a search process is sometimes called an *expanding* search (Leamer 1978). Its chief drawback is that often there are too many possible expansions to consider within a reasonable length of time. If, however, one neglects to consider any possible expansion, one risks missing an important shortcoming of the initial model. For example, if our minimal model involves only single "first-order" terms ("main effects") for 12 variables, we would have $\binom{12}{2}=66$ possible two-way products among these variables to consider, as well as 12 quadratic terms, for a total of 78 possible expansions with just one second-order term. An analyst may not have the time, patience, or resources to examine all the possibilities in detail; this predicament usually leads to use of automatic significance-testing procedures to select additional terms, which (as referenced above) can lead to distorted statistics.

Some strategies begin by specifying an initial model form that is flexible enough to approximate any credible model form. A flexible starting point can be less demanding than a simple one in terms of need for background information.

For example, rather than concern ourselves with what the literature suggests about the shape of a dose-response curve, we can employ a starting model form that can approximate a wide range of curves. Similarly, rather than concern ourselves with what the literature suggests about joint effects, we can employ a form that can approximate a wide range of joint effects. We can then search for a simpler but adequate model by removing from the flexible model any complexities that appear unnecessary in light of the data. Such a search process, based on simplifying a complex model, is sometimes called a *contracting* or simplifying search (Leamer 1978). The same criteria for adding terms can instead be used for deleting terms; again, see Sect. 30.10.3 "Tests of Fit."

The chief drawback of a purely contracting search is that a sufficiently flexible prior model may be too complex to fit to the available data. This is because more complex models generally involve more parameters; with more parameters in a model, more data are needed to produce trustworthy point and interval estimates. Standard model-fitting methods may yield biased estimates or may completely fail to yield any estimates (e.g., not converge) if the fitted model is too complex. For example, if our flexible model for 12 variables contains all first- and second-order terms, there will be 12 first-order plus 12 quadratic plus 66 product terms, for a total of 90 coefficients. Fitting this model may be well beyond what our data or computing resources can support.

Because of potential fitting problems, contracting searches begin with something much less than a fully flexible model. Some begin with a model as flexible as can be fit or maximal model. As with minimal models, maximal models are not unique. In order to produce a model that can be fit, one may have to limit flexibility of dose-response, flexibility of joint effects, or both. It is also possible to start a model search anywhere in between the extremes of minimal and maximal models and proceed by expanding as seems necessary and contracting as seems reasonable based on the data (although again, resource limitations usually lead to mechanical use of significance tests for this process). Unsurprisingly, such *stepwise* searches share some advantages and disadvantages with purely expanding and purely contracting searches. Like other searches, care should be taken to insure that the starting and ending points do not conflict with prior information.

The results obtained from a model search can be very sensitive to the choice of starting model. One can check for this problem by conducting several searches, starting at different models. However, there are always too many possible starting models to check them all. Thus, if one has many variables (and hence many possible models) to consider, model search strategies will always risk producing a misleading model form. A traditional but potentially cumbersome way to address this problem is to present results from a number of different models, which leads into issues of how to select which models to present.

Modern methods to address this problem and other modeling issues can be found in the literature on statistical learning, for example, under topics such as cross-validation, model averaging, semi-parametric and non-parametric regression, smoothing, bootstrapping, algorithmic modeling, and boosting (Harrell 2001; Breiman 2001; Ruppert et al. 2003; Hastie et al. 2009). There is also a large Bayesian literature on the topic (e.g., Brown et al. 2002). Some of these methods reduce the selection problem by merging models or their results, for example, by averaging over competing models (Draper 1995; Raftery 1995; Carlin and Louis 2000; Viallefont et al. 2001; Hastie et al. 2009), by embedding the models in a hierarchical model that contains them all as special cases (Greenland 1999), or by using inferential methods that account for the selection process (Ye 1998; Harrell 2001; Ruppert et al. 2003; Efron 2004; Shen et al. 2004; Hastie et al. 2009).

30.9 Model Fitting

30.9.1 Residual Distributions

Different fitting methods can lead to different estimates; thus, in presenting results, one should specify the method used to derive the estimates. The vast majority of programs for risk and rate modeling use *maximum-likelihood* (ML) estimation, which is based on very specific assumptions about how the observed values of Y tend to distribute (vary) when the vector of regressors $\boldsymbol{X}$ is fixed at a given value $\boldsymbol{x}$. This distribution is called the error distribution or *residual distribution* of Y.

If Y is the person-time rate observed at a given level $\boldsymbol{x}$ of $\boldsymbol{X}$, and T is the corresponding observed person-time, it is conventionally assumed that the number of cases observed, $A = YT$, would tend to vary according to a Poisson distribution if the person-time were fixed at its observed value. Hence, conventional ML regression analysis of person-time rates is usually called *Poisson regression.* If, on the other hand, Y is the proportion of cases observed at a given level $\boldsymbol{x}$ of $\boldsymbol{X}$ out of a person-count total N, it is conventionally assumed that the number of cases observed, $A = YN$, would tend to vary according to a *binomial distribution* if the number of persons (person count) N was fixed at its observed value. Hence, conventional ML regression analysis of prevalence or incidence proportions (average risks) is sometimes called *binomial regression.* Note that if $N = 1$, the proportion diseased Y can be only 0 or 1; in this situation, $A = YN$ can be only 0 or 1 and is said to have a Bernoulli distribution (which is just a binomial distribution with $N = 1$). The binomial distribution can be deduced from the homogeneity and independence assumptions. Its use is inadvisable if there are important violations of either assumption, for example, if the disease is contagious over the study period.

If Y is the number of exposed cases in a 2 × 2 table, the conventionally assumed distribution for Y is the *hypergeometric;* ML fitting in this situation is usually referred to as *conditional maximum likelihood* (CML). CML fitting is the method used for conditional logistic regression; it is closely related to partial-likelihood methods, which are used for fitting Cox models in survival analysis.

Basics of maximum likelihood in epidemiology can be found in Breslow and Day (1980, 1987), Hosmer and Lemeshow (2000), Jewell (2004), and Greenland and Rothman (2008). More advanced treatments of maximum likelihood and its extensions to partial, conditional, penalized, and other types of likelihood can be found in many books, including McCullagh and Nelder (1989), Harrell (2001), and Ruppert et al. (2003).

30.9.2 Overdispersion

What if the residual distribution of the observed Y does *not* follow the conventionally assumed residual distribution? Under a broad range of conditions, it can be shown that the resulting ML fitted values (ML estimates) will remain approximately unbiased if no other source of bias is present (White 1994). Nonetheless, the standard errors obtained from the program will be biased. In particular, if the actual variance of Y given $X = x$ (the *residual variance)* is larger than that implied by the conventional distribution, Y is said to suffer from *overdispersion* or *extravariation,* and the standard errors and p-values obtained from an ordinary maximum-likelihood regression program will be too small.

In Poisson regression, overdispersion is sometimes called "extra-Poisson variation"; in binomial regression, overdispersion is sometimes called "extrabinomial variation." Typically, such overdispersion arises when there is dependence among the recorded outcomes, as when the outcome Y is the number infected in a group, or Y is the number of times a person gets a disease. As an example, suppose Y is the number of eyes affected by glaucoma in an individual. In a natural population, $Y = 0$ for most people and $Y = 2$ for most of the remainder, with $Y = 1$ very infrequently. In other words, the Y values would be largely limited to the extremes of 0 and 2. In contrast, a binomially distributed variable with the same possible values (0, 1, or 2) and the same mean as Y would have a higher probability of 1 than 2 and hence a smaller variance than Y.

Various approaches have been developed to cope with potential overdispersion, most of which are based on modeling the residual distribution. One approach is to use maximum likelihood, but with a residual distribution that allows a broader range of variation for Y, such as the negative binomial in place of the Poisson or the beta-binomial in place of the binomial (McCullagh and Nelder 1989). Such approaches have been implemented in some software. Another, simpler approach is to model only the residual variance of Y, rather than completely specify the residual distribution, as in quasilikelihood or pseudolikelihood estimation (McCullagh and Nelder 1989; Ruppert et al. 2003). An even simpler approach uses robust ("sandwich") variance estimation (Harrell 2001; Ruppert et al. 2003), as is the norm in generalized estimating equation (GEE) methods (Diggle et al. 2002).

30.9.3 Sample-Size Considerations

One drawback of all the above fitting methods is that they depend on "large-sample" (asymptotic) approximations, which usually require that the number of parameters in the model is much less than the number of cases observed; a minimum of 10:1 for the case/parameter ratio is a common empirical criterion (e.g., Pike et al. 1980; Peduzzi et al. 1996), although in some examples, even that appears insufficient (Greenland et al. 2000). When the sample size is insufficient, highly distorted estimates may result.

Methods that do not use large-sample approximations (exact methods) can be used to fit certain models. These methods require the same strong distributional assumptions as maximum-likelihood methods. An example is exact logistic regression (Hirji 2006). Unfortunately, exact fitting methods for incidence and prevalence models have both computational and theoretical demands that have limited their use to only a narrow range of models. Also, they do not address all the problems arising from coefficient instability in small samples, such as failure to constrain the size of estimated relative risks (Greenland et al. 2000; Greenland 2006). Alternative approaches are available that meet these objections while providing approximately valid small-sample results. These alternatives include penalized likelihood estimation and related "shrinkage" methods of Stein estimation, random-coefficient regression, ridge regression, and Bayesian regression (Efron and Morris 1975; Copas 1983; Titterington 1985; Le Cessie and van Houwelingen 1992; Carlin and Louis 2000; Greenland 2000b, 2001, 2007; Harrell 2001; Ruppert et al. 2003).

30.9.4 Data Grouping and Count Data

The usual input to regression programs is in the form of an observed numerator of case counts and a corresponding denominator of counts or person-time, one pair for each observed value $\boldsymbol{x}$ of the vector of regressors $\boldsymbol{X}$. It is commonly thought that rate models such as Eq. 30.17 require the regressors to be grouped into categories. This is not so: Most rate-regression programs can make use of ungrouped regressor data, in which an individual record contains an indicator of whether the individual got disease, which serves as the case count (1 if the individual got disease, 0 if not) and also contains the person-time at risk observed for the person. The only issue is then whether there are sufficient data overall to allow the fitting methods to work properly (Greenland et al. 2000). Conversely, suppose that each component of $\boldsymbol{X}$ is a categorical (discrete) variable, such as marital status, parity, sex, and age category. The multiway contingency table formed by cross-classifying all subjects on each $\boldsymbol{X}$ component then has a cell for each value $\boldsymbol{x}$ of $\boldsymbol{X}$. Such tables are often analyzed using a log-linear model for the number of observations (subject count) A_x expected in cell $\boldsymbol{x}$ of the table, $\ln(A_x) = \alpha + \boldsymbol{x\beta}$. Although software for log-linear count modeling is available in some packages, count data can be easily analyzed using any program for Poisson (rate) regression, by entering A_x as the rate numerator at regressor value $\boldsymbol{x}$, and giving each count a person-time denominator of 1. This approach eliminates the need for special software and provides greater modeling flexibility than is available in many tabular-analysis packages.

30.10 Model Checking

It is important to check a fitted model against the data. The extent of these checks may depend on what purpose we wish the model to serve. At one extreme, we may only wish the fitted model to provide approximately valid *summary* estimates or trends for a few key relationships. For example, we might wish only to estimate the average increment in risk produced by a unit increase in exposure. At the other extreme, we may want the model to provide approximately valid *regressor-specific* predictions of outcomes, such as exposure-specific risks by age, sex, and ethnicity. The latter goal is more demanding and requires more detailed scrutiny of results, sometimes on a subject-by-subject basis.

Model diagnostics can detect discrepancies between data and a model only within the range of the data and then only where there are enough observations to provide adequate diagnostic power. For example, there is much controversy concerning the health effects of low-dose radiation exposure (exposures that are only modestly in excess of natural background levels). This controversy arises because the natural incidence of key outcomes (such as leukemia) is low, and few cases have been observed in low-dose cohorts. As a result, several proposed dose-response models "fit the data adequately" in the low-dose region in that each model passes the standard battery of diagnostic checks. Nonetheless, the health effects predicted by these models conflict to an important extent.

More generally, one should bear in mind that a good-fitting model is not the same as a correct model. In particular, a model may appear correct in the central range of the data, but produce grossly misleading predictions for combinations of covariate values that are poorly represented or absent in the data.

30.10.1 Tabular Checks

Both tabular methods (such as Mantel and Haenszel 1959) and regression methods produce estimates by merging assumptions about population structure (such as that of a common odds ratio or of an explicit regression model) with observed data. When an estimate is derived using a regression model, especially one with many regressors, it may become difficult to judge how much the estimate reflects the data and how much it reflects the model.

To investigate the source of results, we recommend one compare model-based results to the corresponding tabular (categorical-analysis) results. As an illustration, suppose we wish to check a logistic model in which X_1 is the exposure under study, and four other regressors X_2, X_3, X_4, X_5 appear in the model, with X_1, X_2, X_3 continuous, X_4, X_5 binary, and products among X_1, X_2, and X_4 in the model. Any regressor in a model must appear in the corresponding tabular analysis. Because X_2 and X_4 appear in products with X_1 and the model is logistic, they should be treated as modifiers of the X_1 odds ratio in the corresponding tabular analysis. X_3 and

X_5 do not appear in products with X_1 and so should be treated as pure confounders (adjustment variables) in the corresponding tabular analysis. Because X_1, X_2, X_3 are continuous in the model, they must have at least three levels in the tabular analysis so that the results can at least crudely reflect trends seen with the model. If all three of these regressors were categorized into four levels, the resulting table of disease (two levels) by all regressors would have $2 \times 4^3 \times 2^2 = 512$ cells and perhaps many zero cells.

From this table, we would attempt to compute 3 (for exposure strata 1,2,3, versus 0) adjusted odds ratios (e.g., Mantel-Haenszel) for each of the $4 \times 2 = 8$ combinations of X_2 and X_4, adjusting all $3 \times 8 = 24$ odds ratios for the $4 \times 2 = 8$ pure-confounder levels. Some of these 24 adjusted odds ratios might be infinite or undefined due to small numbers, which would indicate that the corresponding regression estimates are largely model projections. Similarly, the tabular estimates might not exhibit a pattern seen in the regression estimates, which would suggest that the pattern was induced by the regression model rather than the data. For example, the regression estimates might exhibit a monotone trend with increasing exposure even if the tabular estimates did not. Interpretation of such a conflict would depend on the context: If we were certain that dose-response was monotone (e.g., smoking and esophageal cancer), the monotonicity of the regression estimates would favor their use over the tabular results; in contrast, doubts about monotonicity (e.g., as with alcohol and coronary heart disease) would lead us to use the tabular results or search for a model that did not impose monotonicity.

30.10.2 Tests of Regression and R^2

Most programs supply a "test of regression" or "test of model," which is a test of the hypothesis that all the regression coefficients (except the intercept α) are zero. For instance, in the exponential rate model

$$I(\boldsymbol{x}) = \exp(\alpha + \boldsymbol{x}\boldsymbol{\beta}),$$

the "test of regression" provides a p-value for the null hypothesis that all the components of $\boldsymbol{\beta}$ are zero, that is, that $\beta_1 = \ldots = \beta_n = 0$. Similarly, the "test of R^2" provided by linear regression programs is just a test that all the regressor coefficients are zero. A small p-value from these tests suggests that the variation in outcomes observed across regressor values appears improbably large under the hypothesis that the regressors are unrelated to the outcome. Such a result suggests that at least one of the regressors is related to the outcome. It does *not,* however, imply that the model fits well or is adequate in any way.

To understand the latter point, suppose that $\boldsymbol{X}$ comprises the single indicator $X_1 = 1$ for smokers, 0 for non-smokers, and the outcome Y is average year risk of lung cancer. In most any study of reasonable size and validity, "the test of regression" (which here is just a test of $\beta_1 = 0$) would yield a small p-value. Nonetheless, the model would be inadequate to describe variation in risk because

Table 30.1 Hypothetical cohort data illustrating inappropriateness of R^2 for binary outcomes (see text)

	$X = 1$	$X = 0$
$Y = 1$	1,900	100
Total	100,000	100,000

Risk ratio = 19, $R^2 = 0.008$

it neglects amount smoked, age at start, and sex. More generally, a small p-value from the test of regression only tells us that at least one of the regressors in the model should be included in some form or another; it does not tell us which regressor or what form to use, nor does it tell us anything about what was left out of the model. Conversely, a large p-value from the "test of regression" does not imply that all the regressors in the model are unimportant or that the model fits well. It is always possible that transformations of those regressors would result in a small p-value, or that their importance cannot be discerned given the random error in the data.

A closely related mistake is interpreting the squared multiple-correlation coefficient R^2 for a regression as a goodness-of-fit measure. R^2 only indicates the proportion of Y variance that is attributable to variation in the fitted mean of Y. While $R^2 = 1$ (the largest possible value) does correspond to a perfect fit, R^2 can also be close to zero under a correct model if the residual variance of Y (i.e., the variance of Y around the true regression curve) is always close to the total variance of Y.

The preceding limitations of R^2 apply in general. Correlational measures such as R^2 can become patently absurd measures of fit or association when the regressors and regressand are discrete or bounded (Rosenthal and Rubin 1979; Greenland et al. 1986; Cox and Wermuth 1992; Greenland 1996). As an example, consider Table 30.1 showing a large association of a factor with a rare disease. The logistic model $R(x) = \text{expit}(\alpha + \beta x)$ fits these data perfectly because it uses two parameters to describe only two proportions. Furthermore, $X = 1$ is associated with a 19-fold increase in risk. Yet the correlation coefficient for X and Y (derived using standard formulas) is only 0.09, and the R^2 for the regression is only 0.008.

Correlation coefficients and R^2 can give even more distorted impressions when multiple regressors are present (Greenland et al. 1986, 1991). For this reason, we strongly recommend against their use as measures of association or effect when modeling incidence or prevalence.

30.10.3 Tests of Fit

Tests of model fit check for non-random incompatibilities between the fitted regression model and the data. To do so, however, these tests must assume that the fitting method used was appropriate; in particular, test validity may be sensitive to assumptions about the residual distribution that were used in fitting. Conversely, it is possible to test assumptions about the residual distribution, but these tests usually have little power to detect violations unless a parametric regression model is assumed. Thus, useful model tests cannot be performed without making some assumptions.

Many tests of regression models are *relative* in that they test the fit of an index model by assuming the validity of a more elaborate *reference* model that contains it. A test that assumes a relatively simple reference model (i.e., one that has only a few more coefficients than the index model) will tend to have better power than a test that assumes a more complex reference model, although it will be valid only under narrower conditions.

When models are fit by maximum likelihood (ML), a standard method for testing the fit of a simpler model against a more complex model is the *deviance test,* also known as the likelihood-ratio test. Suppose that X_1 represents cumulative dose of an exposure and that the index model we wish to test is

$$R(x_1) = \text{expit}(\alpha + \beta_1 x_1)$$

a simple linear-logistic model. When we fit this model, an ML program should supply either a "residual deviance statistic" $D(\tilde{\alpha}, \tilde{\beta}_1)$ or a "model log-likelihood" $L(\tilde{\alpha}, \tilde{\beta}_1)$, where $\tilde{\alpha}, \tilde{\beta}_1$ are the ML estimates for this simple model. Suppose we wish to test the fit of the index model taking as the reference the fractional-polynomial logistic model

$$R(x_1) = expit(\alpha + \beta_1 x_1 + \beta_2 x_1^{1/2} + \beta_3 x_1^2).$$

We then fit this model and get either the residual deviance $D(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ or the log-likelihood $L(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ for the model, where $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ are the ML estimates for this power model. The deviance statistic for testing the linear-logistic model against the power-logistic model (i.e., for testing $\beta_2 = \beta_3 = 0$) is then

$$\Delta D(\beta_2, \beta_3) = D(\tilde{\alpha}, \tilde{\beta}_1) - D(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3).$$

This statistic is related to the model log-likelihoods by the equation

$$\Delta D(\beta_2, \beta_3) = -2[L(\tilde{\alpha}, \tilde{\beta}_1) - L(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)]$$

(McCullagh and Nelder 1989). If the linear-logistic model is correct (so that $\beta_2 = \beta_3 = 0$) and the sample is large enough, this statistic has an approximate χ^2 distribution with 2 degrees of freedom, which is the difference in the number of parameters in the two models.

A small p-value from this statistic suggests that the linear-logistic model is inadequate or fits poorly; in some way, either or both the terms $\beta_2 x_1^{1/2}$ and $\beta_3 x_1^2$ capture deviations of the true regression from the linear-logistic model. A large p-value does *not,* however, imply that the linear-logistic model is adequate or fits well; it means only that no need for the terms $\beta_2 x_1^{1/2}$ and $\beta_3 x_1^2$ was detected by the test. In particular, a large p-value from this test leaves open the possibility that $\beta_2 x_1^{1/2}$ and $\beta_3 x_1^2$ are important for describing the true regression function, but the test failed to detect this condition; it also leaves open the possibility that some other terms not present in the reference model may be important in the same sense. These unexamined terms may involve X_1 or other regressors.

Now consider a more general description. Suppose that we wish to test an index model against a reference model in which it is nested (contained) and that this reference model contains p more unknown parameters (coefficients) than the index model. We fit both models and obtain either residual deviances of D_i and D_r for the index and reference models or log-likelihoods L_i and L_r. If the sample is large enough and the index model is correct, the deviance statistic

$$\Delta D = D_i - D_r = -2(L_i - L_r) \tag{30.56}$$

will have an approximate χ^2 distribution with p degrees of freedom. Again, a small p-value suggests that the index model does not fit well, but a large p-value does *not* mean the index model fits well, except in the very narrow sense that the test did not detect a need for the extra terms in the reference model.

Whatever the size of the deviance p-value, its validity depends on three assumptions (in addition to absence of the usual biases). First, it assumes that ML fitting of the models is appropriate; in particular, there must be enough subjects to justify use of ML to fit the reference model, and the assumed residual distribution must be correct. Second, it assumes that the reference regression model is approximately correct. Third, it assumes that the index model being tested is nested within the reference model. The third is the only assumption that is easy to check: In the previous example, we can see that the linear-logistic model is just the special case of the power-logistic model in which $\beta_2 = \beta_3 = 0$. In contrast, if we used the linear-logistic model as the index model (as above) but used the power-linear model

$$R(x_1) = \alpha + \beta_1 x_1 + \beta_2 x_1^{1/2} + \beta_3 x_1^2$$

as the reference model, the resulting deviance difference would be meaningless because the latter model does *not* contain the linear-logistic model as a special case.

Comparison of non-nested models is a more difficult task unless the compared models have the same number of parameters. Criteria that have been commonly suggested for comparison purposes include the *Akaike information criterion* (AIC), often defined as the model deviance plus $2p$, where p is the number of model parameters; the Schwarz information criterion or *Bayesian information criterion* (BIC), often defined as the model deviance plus $p \cdot \ln(N)$, where N is the number of observations; criteria based on *cross-validation*; and criteria based on *Bayes factors*. Larger values of these criteria suggest poorer out-of-sample predictive performance of the model. The definitions of AIC and BIC vary considerably across textbooks; fortunately, the differences do not affect the ranking of models by each criterion. For details on these criteria, see Harrell (2001), Ruppert et al. (2003), or Hastie et al. (2009).

30.10.4 Global Tests of Fit

One special type of deviance test of fit can be performed when Y is a proportion or rate. Suppose that, for every distinct regressor level $\boldsymbol{x}$, at least four cases would be expected if the index model were correct; also, if Y is a proportion, suppose at least four non-cases would be expected if the index model were correct. (This criterion, while somewhat arbitrary, originated because it ensures that the chance of a cell count being zero is less than 2% if the cell variation is Poisson and the index model is correct.) We can then test our index model against the *saturated* regression model

$$E(Y|\boldsymbol{X} = \boldsymbol{x}) = \alpha_x,$$

where α_x is a distinct parameter for every distinct observed level $\boldsymbol{x}$ of $\boldsymbol{X}$; that is, α_x may represent a different number for every level of $\boldsymbol{X}$ and may vary in any fashion as $\boldsymbol{X}$ varies. This model is so general that it contains all other regression models as special cases.

The degrees of freedom for the test of the index model against the saturated model is the number of distinct $\boldsymbol{X}$-levels (which is the number of parameters in the saturated model) minus the number of parameters in the index model and is often called the *residual degrees of freedom* for the model. This *residual* deviance test is sometimes called a "global test of fit" because it has some power to detect any systematic incompatibility between the index model and the data. Another well-known global test of fit is the *Pearson* χ^2 *test,* which has the same degrees of freedom and sample-size requirements as the saturated-model deviance test.

Suppose we observe K distinct regressor values and we list them in some order, $x_1, \ldots, x_K$. The statistic used for the Pearson test has the form of a residual sum of squares:

$$RSS_{Pearson} = \sum\nolimits_k (Y_k - \hat{Y}_k)^2/\hat{V}_k = \sum\nolimits_k [(Y_k - \hat{Y}_k)/\hat{S}_k]^2,$$

where the sum is over all observed values $1, \ldots, K$, Y_k is the rate or risk observed at level $\boldsymbol{x}_k$, $\hat{Y}_k$ is the rate or risk predicted (fitted) at x_k by the model, $\hat{V}_k$ is the estimated variance of $\hat{Y}_k$ when $\boldsymbol{X} = \boldsymbol{x}_k$, and $\hat{S}_k = \hat{V}_k^{1/2}$ is the estimated standard deviation of Y_k under the model. In Poisson regression, $\hat{Y}_k = \exp(\hat{\alpha} + \boldsymbol{x}_k\hat{\boldsymbol{\beta}})$ and $\hat{V}_k = \hat{Y}_k/T_k$, where T_k is the person-time observed at x_k; in binomial regression, $\hat{Y}_k = \text{expit}(\hat{\alpha} + \boldsymbol{x}_k\hat{\boldsymbol{\beta}})$ and $\hat{V}_k = \hat{Y}_k(1 - \hat{Y}_k)/N_k$, where N_k is the number of persons observed at $\boldsymbol{x}_k$. The quantity $(Y_k - \hat{Y}_k)/\hat{S}_k$ is sometimes called the *standardized residual* at level $\boldsymbol{x}_k$; it is the distance between Y_k and $\hat{Y}_k$ expressed in units of the estimated standard deviation of Y_k under the model.

Other global tests have been proposed that have fewer degrees of freedom and less restrictive sample-size requirements than the deviance and Pearson tests (Hosmer and Lemeshow 2000). A major drawback of all global tests of fit, however, is their low power to detect model problems (Hosmer et al. 1997). If any of the tests yields a low p-value, we can be confident the tested (index) model fails to fit in some fashion and thus may need modification or replacement, depending on the nature of the failure (albeit the tests provide no clue as to how to proceed).

On the one hand, if the only purpose of the model is to estimate and compare average risks over broad subgroups, as in model-based standardization (Greenland 2004, 2008a), only very large or qualitative discrepancies may be of concern. On the other hand, if prediction of individual outcomes is the goal, as in clinical applications (Steyerberg 2009), a high p-value from all global tests does not mean the model is satisfactory. In fact, the tests are unlikely to detect any but very large conflicts between the fitted model and the data. Therefore, global tests should be regarded as crude preliminary screening tests only to allow quick rejection of grossly unsatisfactory models.

The deviance and Pearson statistics are sometimes used directly as measures of distance between the data and the model. Such use is most easily seen for the Pearson statistic. The second form of the Pearson statistic shows that it is the sum of squared standardized residuals; in other words, it is a sum of squared distances between data values and model-fitted values of Y. The deviance and Pearson global test statistics can also be transformed into measures of prediction error under the model (McCullagh and Nelder 1989; Hosmer and Lemeshow 2000).

30.10.5 Model Diagnostics

Suppose now we have found a model that has passed preliminary checks such as tests for additional terms and global tests of fit. Before adopting this model as a source of estimates, it is wise to further check the model against the basic data and assess the trustworthiness of any model-based inferences we wish to draw. Such activity is subsumed under the topic of *model diagnostics* and its subsidiary topics of residual analysis, influence analysis, and model sensitivity analysis. These topics are vast, and we can only mention a few approaches here. In particular, we neglect the classical topic of residual analysis, largely because its proper usage involves a number of technical complexities when dealing with the censored data and non-linear models predominant in epidemiology. Detailed treatments of diagnostics for such models can be found in Hosmer and Lemeshow (2000), Harrell (2001), and Ruppert et al. (2003).

30.10.6 Delta-Beta Analysis

One important and simple diagnostic tool available in some packaged software is *delta-beta* ($\Delta\beta$) *analysis.* For a data set with N subjects total, estimated model coefficients (or approximations to them) are recomputed N times over, each time deleting exactly one of the subjects from the model fitting. Alternatively, for individually matched data comprising N matched sets, the delta-beta analysis may be done deleting one set at a time. In either approach, the output is N different sets of coefficients estimates: These sets are then examined to see if anyone subject or matched set influences the resulting estimates to an unusual extent.

To illustrate, suppose our objective is to estimate the rate ratio per unit increase in an exposure X_1, to be measured by $\exp(\hat{\beta}_1)$, where $\hat{\beta}_1$ is the estimated exposure coefficient in an exponential rate model. For each subject, the entire model (confounders included) is re-fit without that subject. Let $\hat{\beta}_{1(-i)}$ be the estimate of $\hat{\beta}_1$ obtained when subject i is excluded from the data. The difference $\hat{\beta}_{1(-i)} - \hat{\beta}_1 \equiv \Delta\hat{\beta}_{1(-i)}$ is called the *delta-beta* for β_1 for subject i. The influence of subject i on the results can be assessed in several ways. One way is to examine the impact on the rate-ratio estimate. The proportionate change in the estimate from dropping subject i is

$$\exp(\hat{\beta}_{1(-i)}) / \exp(\hat{\beta}_1) = \exp(\hat{\beta}_{1(-i)} - \hat{\beta}1) = \exp(\Delta\hat{\beta}_{1(-i)}),$$

for which a value of 1.30 indicates dropping subject i increases the estimate by 30%, and a value of 0.90 indicates dropping subject i decreases the estimate by 10%. One can also assess the impact of dropping the subject on confidence limits, p-values, or any other quantity of interest.

Some packages compute "standardized" delta-betas, $\Delta\hat{\beta}_{1(-i)}/\hat{s}_1$ where $\hat{s}_1$ is the estimated standard deviation for $\hat{\beta}_1$. By analogy with Z-statistics, any standardized delta-beta below -1.96 or above 1.96 is sometimes interpreted as being unusual. This interpretation can be misleading, however, because the standard deviation used in the denominator is not that of the delta-beta. A standardized delta-beta is only a measure of the influence of an observation expressed in units of standard error.

It is possible that one or a few subjects or matched sets are so influential that deleting them alters the conclusions of the study, even when N is in the hundreds (Pregibon 1981). In such situations, comparison of the records of those subjects to others may reveal unusual combinations of regressor values among those subjects. Such unusual combinations may arise from previously undetected data errors and should at least lead to enhanced caution in interpretation. For instance, it may be only mildly unusual to see a woman who reports having had a child at age 45 or a woman who reports natural menopause at age 45. The combination in one subject, however, may arouse suspicion of a data error in one or both regressors, a suspicion worth the labor of further data scrutiny if that woman or her matched set disproportionately influences the results.

Delta-beta analysis must be replaced by a more complex analysis if the exposure of interest appears in multiple model terms, such as indicator terms, power terms, product terms, or spline terms. In that situation, one must focus on changes in estimates of specific effects or summaries, for example, changes in estimated risk ratios.

30.10.7 Model Sensitivity Analysis

A valuable diagnostic procedure, which is often a by-product of the model-selection approaches described earlier, is model sensitivity analysis (Draper 1995;

Saltelli et al. 2000). Given the variety of models available for fitting, it is inevitable that quite a few can be found that fit reasonably well by all standard tests and diagnostics. Model sensitivity analysis seeks to identify a spectrum of such acceptable models and asks whether various estimates and tests are sensitive to the choice of model used for inference. On the one hand, those results that are consistent (stable) across acceptable model-based and stratified analyses can be presented using any one of the analyses. On the other hand, results that vary across analyses need to be reported with much more uncertainty than is indicated by their (unstable) confidence intervals; in particular, one should report the fact that the estimates were unstable.

The credibility, value, and results of a sensitivity analysis are themselves sensitive to the spectrum of models investigated for acceptability. Most such analyses cover limited territory, in part because most software has a restricted range of model forms that can be examined easily. Typically, only exponential-Poisson and Cox models for rates and binomial-logistic models for risks are available, although some packages supply linear rate and linear odds models. The similar restrictions imposed by these models result in severe limitations on the range of analysis. Nonetheless, within these limits, there are vast possibilities for the terms representing the effects of exposures, confounders, and modifiers.

30.11 Conclusions

The topic of regression modeling is vast and even an extended chapter such as the present one can provide only brief overviews of its key concepts and tools. This present chapter has focused on concepts and model interpretation rather than model fitting and computational methods in the hopes of providing a foundation for proper use of those methods. For more detailed coverage of concepts and methods in the general regression context, see the cited textbooks, in particular Berk (2004) and Harrell (2001).

References

Agresti A (2002) Categorical data analysis. Wiley, New York

Ananth CV, Kleinbaum DG (1997) Regression models for ordinal responses: a review of methods and applications. Int J Epidemiol 26:1323–1333

Bancroft TA, Han C-P (1977) Inference based on conditional specification: a note and a bibliography. Int Stat Rev 45:117–127

Berk R (2004) Regression analysis: a constructive critique. Sage publications, Thousand Oaks

Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT Press, Cambridge

Breiman L (2001) Statistical modeling: the two cultures (with discussion). Stat Sci 16:199–231

Breslow NE, Day NE (1980) Statistical methods in cancer research. Vol I: the analysis of case-control data. IARC, Lyon

Breslow NE, Day NE (1987) Statistical methods in cancer research. Vol II: the design and analysis of cohort studies. IARC, Lyon

Brown PJ, Vannucci M, Fearn T (2002) Bayes model averaging with selection of regressors. J R Stat Soc Ser B 64:519–536

Carlin B, Louis TA (2000) Bayes and empirical-Bayes methods of data analysis, 2nd edn. Chapman and Hall, New York

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu C (2006) Measurement error in nonlinear models, 2nd edn. Chapman and Hall, New York

Cole SR, Ananth CV (2001) Regression models for unconstrained, partially or fully constrained continuation odds ratios. Int J Epidemiol 30:1379–1382

Copas JB (1983) Regression, prediction, and shrinkage (with discussion). J R Stat Soc B 45: 311–354

Cox DR (1972) Regression models and life tables (with discussion). J R Stat Soc B 34:187–220

Cox DR, Oakes D (1984) Analysis of survival data. Chapman and Hall, New York

Cox DR, Wermuth N (1992) A comment on the coefficient of determination for binary responses. Am Stat 46:1–4

Diggle PJ, Heagerty P, Liang KY, Zeger SL (2002) The analysis of longitudinal data, 2nd edn. Oxford University Press, New York

Draper D (1995) Assessment and propagation of model uncertainty. J R Stat Soc Ser B 57:45–97

Draper NR, Guttman I, Lapczak L (1979) Actual rejection levels in a certain stepwise test. Commun Stat A 8:99–105

Easton DF, Peto J, Babiker AG (1991) Floating absolute risk: an alternative to relative risk in survival and case-control analysis avoiding an arbitrary reference group. Stat Med 10:1025–1035

Efron B (2004) The estimation of prediction error: covariance penalties and cross-validation. J Am Stat Assoc 99:619–642

Efron B, Morris CN (1975) Data analysis using Stein's estimator and its generalizations. J Am Stat Assoc 70:311–319

Faraway JJ (1992) On the cost of data analysis. J Comput Graph Stat 1:213–219

Flack VF, Chang PC (1987) Frequency of selecting noise variables in subset regression analysis: a simulation study. Am Stat 41:84–86

Freedman DA (1983) A note on screening regression equations. Am Stat 37:152–155

Freedman DA, Navidi W, Peters SC (1988) On the impact of variable selection in fitting regression equations. In: Dijlestra TK (ed) On model uncertainty and its statistical implications. Springer, Berlin, pp 1–16

Greenland S (1993) Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary testing, and empirical-Bayes regression. Stat Med 12:717–736

Greenland S (1994) Alternative models for ordinal logistic regression. Stat Med 13:1665–1677

Greenland S (1995a) Dose-response and trend analysis: alternatives to categorical analysis. Epidemiology 6:356–365

Greenland S (1995b) Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. Epidemiology 6:450–454

Greenland S (1995c) Problems in the average-risk interpretation of categorical dose-response analyses. Epidemiology 6:563–565

Greenland S (1996) A lower bound for the correlation of exponentiated bivariate normal pairs. Am Stat 50:163–164

Greenland S (1999) Multilevel modeling and model averaging. Scand J Work Environ Health 25(suppl 4):43–48

Greenland S (2000a) Principles of multilevel modeling. Int J Epidemiol 29:158–167

Greenland S (2000b) When should epidemiologic regressions use random coefficients? Biometrics 56:915–921

Greenland S (2001) Putting background information about relative risks into conjugate priors. Biometrics 57:663–670

Greenland S (2003) The impact of prior distributions for uncontrolled confounding and response bias: a case study of the relation of wire codes and magnetic fields to childhood leukemia. J Am Stat Assoc 98:47–54

Greenland S (2004) Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. Am J Epidemiol 160:301–305

Greenland S (2005a) Epidemiologic measures and policy formulation: Lessons from potential outcomes (with discussion). Emerg Themes Epidemiol 2:1–4

Greenland S (2005b) Multiple-bias modeling for observational studies. J R Stat Soc Ser A 168:267–308

Greenland S (2006) Bayesian perspectives for epidemiologic research. I. Foundations and basic methods (with comment and reply). Int J Epidemiol 35:765–778

Greenland S (2007) Bayesian perspectives for epidemiologic research. II. Regression analysis. Int J Epidemiol 36:195–202

Greenland S (2008a) Introduction to regression modeling. Chap. 21. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 2nd edn. Lippincott Williams & Wilkins, Philadelphia

Greenland S (2008b) Variable selection and shrinkage in the control of multiple confounders. Am J Epidemiol 167:523–529, Erratum: p 1142

Greenland S (2009a) Bayesian perspectives for epidemiologic research. III. Bias analysis via missing-data methods. Int J Epidemiol 38:1662–1673

Greenland S (2009b). Relaxation penalties and priors for plausible modeling of nonidentified bias sources. Stat Sci 24:195–210

Greenland S, Lash TL (2008) Bias analysis. Chap. 19. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 2nd edn. Lippincott Williams & Wilkins, Philadelphia

Greenland S, Maldonado G (1994). The interpretation of multiplicative model parameters as standardized parameters. *Statistics in Medicine* 13:989–999

Greenland S, Poole C (1995) Interpretation and analysis of differential exposure variability and zero-dose categories for continuous exposures. Epidemiology 6:326–328

Greenland S, Rothman KJ (2008) Fundamentals of epidemiologic data analysis. Chap. 13. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 2nd edn. Lippincott Williams & Wilkins, Philadelphia

Greenland S, Schlesselman JJ, Criqui MH (1986) The fallacy of employing standardized regression coefficients and correlations as measures of effect. Am J Epidemiol 123:203–208

Greenland S, Maclure M, Schlesselman JJ, Poole C, Morgenstern H (1991) Standardized regression coefficients: a further critique and review of some alternatives. Epidemiology 2:387–392

Greenland S, Michels KB, Robins JM, Poole C, Willett WC (1999) Presenting statistical uncertainty in trends and dose-response relations. Am J Epidemiol 149:1077–1086

Greenland S, Schwartbaum JA, Finkle WD (2000) Problems from small samples and sparse data in conditional logistic regression. Am J Epidemiol 151:531–539

Greenland S, Rothman KJ, Lash TL (2008) Concepts of interaction. Chap. 5. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 2nd edn. Lippincott Williams & Wilkins, Philadelphia

Gustafson P (2003) Measurement error and misclassification in statistics and epidemiology. Chapman and Hall, Boca Raton

Gustafson P (2005) On model expansion, model contraction, identifiability, and prior information (with discussion). Stat Sci 20:111–140

Harrell F (2001) Regression modeling strategies. Springer, New York

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York

Hernán MA (2005) Hypothetical interventions to define causal effects—afterthought or prerequisite? Am J Epidemiol 162:618–620

Hirji K (2006) Exact analysis of discrete data. CRC Press/Chapman and Hall, Boca Raton

Hosmer DW, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley, New York

Hosmer DW, Hosmer T, LeCessie S, Lemeshow S (1997) A comparison of goodness-of-fit tests for the logistic regression model. Stat Med 16:965–980

Hurvich DM, Tsai CL (1990) The impact of model selection on inference in linear regression. Am Stat 44:214–217

Izenman AJ (2008) Modern multivariate statistical techniques: regression, classification, and manifold learning. Springer, New York

Jewell NP (2004) Statistics for epidemiology. Chapman and Hall, New York

Lagakos SW (1988) Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. Stat Med 7:257–274

Lash TL, Fox MP, Fink AK (2009) Applying quantitative bias analysis to epidemiologic data. Springer, New York

Le Cessie S, van Houwelingen HC (1992) Ridge estimators in logistic regression. Appl Stat 41:191–201

Leamer EE (1978) Specification searches: ad hoc inference with nonexperimental data. Wiley, New York

Maclure M (1993) Demonstration of deductive meta-analysis: ethanol intake and risk of myocardial infarction. Epidemiol Rev 15:328–351

Maclure M, Greenland S (1992) Tests for trend and dose-response: misinterpretations and alternatives. Am J Epidemiol 135:96–104

Maldonado G, Greenland S (1993a) Interpreting model coefficients when the true model form is unknown. Epidemiology 4:310–318

Maldonado G, Greenland S (1993b) Simulation study of confounder-selection strategies. Am J Epidemiol 138:923–936

Maldonado G, Greenland S (1994) A comparison of the performance of model-based confidence intervals when the correct model form is unknown: coverage of asymptotic means. Epidemiology 5:171–182

Mantel N, Haenszel WH (1959) Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 22:719–748

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman and Hall, New York

Michels KB, Greenland S, Rosner BA (1998) Does body mass index adequately capture the relation of body composition and body size to health outcomes? Am J Epidemiol 147:167–172

Moolgavkar SH, Venzon DJ (1987) General relative risk regression models for epidemiologic studies. Am J Epidemiol 126:949–961

Pearl J (2009) Causality, 2nd edn. Cambridge, New York

Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996) A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 49:1373–1379

Pike MC, Hill AP, Smith PG (1980) Bias and efficiency in logistic analyses of stratified case-control studies. Int J Epidemiol 9:89–95

Pregibon D (1981) Logistic regression diagnostics. Ann Stat 9:705–724

Raftery AE (1995) Bayesian model selection in social research (with discussion). Sociol Methodol 25:111–196

Robins JM, Greenland S (1986) The role of model selection in causal inference from nonexperimental data. Am J Epidemiol 123:392–402

Robins JM, Greenland S (1994) Adjusting for differential rates of prophylaxis therapy for PCP in high- versus low-dose AZT treatment arms in an AIDS randomized trial. J Am Stat Assoc 89:737–749

Robins JM, Blevins D, Ritter G, Wulfsohn M (1992) G-estimation of the effect of prophylaxis therapy for *Pneumocystis carinii* pneumonia on the survival of AIDS patients. Epidemiology 3:319–336. Errata: Epidemiology 1993; 4:189

Robins JM, Greenland S, Hu FC (1999) Estimation of the causal effect of time-varying exposure on the marginal means of a repeated binary outcome. J Am Stat Assoc 94:687–712

Robins JM, Hernán MA, Brumback B (2000) Marginal structural models and causal inference in epidemiology. Epidemiology 11:561–570

Rosenthal R, Rubin DB (1979) A note on percent variance explained as a measure of importance of effects. J Appl Psychol 9:395–396

Rothman KJ, Greenland S, Lash TL (2008) Modern epidemiology, 3rd edn. Lippincott Wolters Kluwer, Philadelphia

Royston P, Sauerbrei W (2008) Multivariable model building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Wiley, New York

Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression. Cambridge, New York

Saltelli A, Chan K, Scott EM (eds) (2000) Sensitivity analysis. Wiley, New York

Sclove SL, Morris C, Radhakrishna R (1972) Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. Ann Math Stat 43:1481–1490

Sheehe P (1962) Dynamic risk analysis in retrospective matched-pair studies of disease. Biometrics 18:323–341

Shen X, Huang H, Ye J (2004) Inference after model selection. J Am Stat Assoc 99:751–762

Steyerberg EW (2009) Clinical prediction models. Springer, New York

Strömberg U (1996) Collapsing ordered outcome categories: a note of concern. Am J Epidemiol 144:421–424

Titterington DM (1985) Common structure of smoothing techniques in statistics. Int Stat Rev 53:141–170

Viallefont V, Raftery AE, Richardson S (2001) Variable selection and Bayesian model averaging in epidemiological case-control studies. Stat Med 20:3215–3230

Weiss RE (1995) The influence of variable selection: a Bayesian diagnostic perspective. J Am Stat Assoc 90:619–625

White H (1994) Estimation, inference, and specification analysis. Cambridge University Press, New York

Ye J (1998) On measuring and correcting the effects of data mining and model selection. J Am Stat Assoc 93:120–131

Bayesian Methods in Epidemiology 31

Leonhard Held

Contents

L. Held
Division of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich, Zurich, Switzerland

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_57,

31.1 Introduction

This chapter gives a self-contained introduction to the Bayesian approach to statistical inference. Standard epidemiological problems such as diagnostic tests, the analysis of prevalence, case-control, and cohort data will serve as examples. More advanced topics, such as empirical Bayes methods and Markov chain Monte Carlo techniques, are also covered.

The Bayesian approach is easy to understand, if the reader is able to follow the actual calculations. Only some basic knowledge of the rules of probability and calculus is needed. A reader, willing to dive into these fairly simple technicalities, will be able to fully appreciate the beauty and simplicity of the Bayesian approach. An appendix summarizes the required technical background.

Modern Bayesian statistics is often performed using so-called Monte Carlo methods based on random numbers simulated on a computer. Many quantities of interest can be computed very easily using Monte Carlo. From time to time, I will show very short programming code in `R` to illustrate the simplicity of Monte Carlo methods. Understanding of the code is, however, not necessary for an understanding of this chapter.

To understand Bayesian methods, in particular Bayes' theorem, the most important concept is that of a *conditional probability*. In Sect. 31.2, we will illustrate the notion of conditional probabilities and Bayesian updating in the context of diagnostic testing. Further details on conditional probabilities are listed in Appendix A. The Bayesian approach to parameter estimation is discussed in Sect. 31.3. Appendix B summarizes important properties of the distributions used in this section and their implementation in `R`. After a brief introduction to Bayesian prediction, Sect. 31.4 discusses techniques for prior criticism and Bayesian model selection. Empirical Bayes methods and Markov chain Monte Carlo techniques are described in Sect. 31.5. We close with some discussion in Sect. 31.6.

31.2 Conditional Probabilities and Diagnostic Testing

The use of Bayes' theorem is routine in the context of diagnostic testing.

Example 31.1. Diagnostic testing
Suppose a simple diagnostic test for a specific disease, which produces either a positive or a negative test result, is known to have 90% sensitivity. This means that the probability of a positive test result, if the person being tested has the disease, is 90%. This is a *conditional* probability since we know that the person has the disease and we write $\Pr(T+ \mid D+) = 0.9$, the probability (Pr) of a positive test result $(T+)$ *given* disease $(D+)$ is 0.9. Now, assume that the test also has 90% specificity and write $D-$ if the person being tested is free of the disease. Similarly, let $T-$ denote a negative test result. A 90% specificity simply means that $\Pr(T- \mid D-) = 0.9$.

Conditional probabilities behave just like ordinary probabilities if we always condition on the same event, for example, on $D+$, say. In particular, they must be numbers between 0 and 1 and $\Pr(T- \mid D+)$ must be equal to $1 - \Pr(T+ \mid D+)$, that is, the conditional probability of a negative test result is 1 minus the conditional probability of a positive test result, if both probabilities are conditional on $D+$. The same of course holds if we condition on $D-$ rather than on $D+$.

However, the real power of conditional probabilities emerges if we condition on different events and relate conditional to ordinary (unconditional) probabilities. The most important formula to do this is *Bayes' theorem* (see Appendix A for a derivation). In the diagnostic testing context, we can use Bayes' theorem to compute the conditional probability of disease given a positive test result:

$$\Pr(D+ \mid T+) = \frac{\Pr(T+ \mid D+)\Pr(D+)}{\Pr(T+)}. \tag{31.1}$$

The prevalence $\Pr(D+)$ is an example of an ordinary (unconditional) probability.

The denominator $\Pr(T+)$ in (31.1), the (unconditional) probability of a positive test result, is unknown, but we know from above that $\Pr(D+ \mid T+)$ and

$$\Pr(D- \mid T+) = \frac{\Pr(T+ \mid D-)\Pr(D-)}{\Pr(T+)} \tag{31.2}$$

must add to unity, from which we can easily deduce that

$$\Pr(T+) = \Pr(T+ \mid D+)\Pr(D+) + \Pr(T+ \mid D-)\Pr(D-). \tag{31.3}$$

This equation is sometimes called the *law of total probability*. Thus, we can calculate $\Pr(T+)$ if we know the sensitivity $\Pr(T+ \mid D+)$, the prevalence $\Pr(D+)$, and $\Pr(T+ \mid D-)$, which is 1 minus the specificity $\Pr(T- \mid D-)$.

Equation 31.1 exemplifies the process of *Bayesian updating*: We update the prior risk $\Pr(D+)$ in the light of a positive test result $T+$ to obtain the posterior risk $\Pr(D+ \mid T+)$, the conditional probability of disease given a positive test result, also known as the *positive predictive value*.

Example 31.1. (Continued)
In the following we assume that the prevalence $\Pr(D+) = 1\%$ for the disease considered above. Then

$$\Pr(T+) = 0.9 \cdot 0.01 + 0.1 \cdot 0.99 = 0.108$$

and hence

$$\Pr(D+ \mid T+) = \frac{0.9 \cdot 0.01}{0.108} \approx 0.083,$$

i.e. the disease risk increases from 1% to 8.3% after a positive test result. It is up to the reader to write down the corresponding formula to (31.1) to compute the *negative predictive value* $\Pr(D- \mid T-)$, which turns out to be approximately 0.999. Thus, the disease risk decreases from 1% to $\Pr(D+ \mid T-)=1 - \Pr(D- \mid T-) = 100\% - 99.9\% = 0.1\%$ if the test was negative. The disease risk changes in the expected direction depending on the test result.

Equation 31.1, with the denominator $\Pr(T+)$ replaced by (31.3), is often referred to as Bayes' theorem in probability textbooks. However, the resulting formula is somewhat complex and not particularly intuitive. A simpler version of

Bayes' theorem can be obtained if we switch from probabilities to *odds*. In general, whenever we take a probability, and divide it by 1 minus that probability, the resulting ratio is referred to as the corresponding odds. Of course, every probability refers to a particular event happening and 1 minus that probability is the probability that the event is *not* happening. Odds are hence nothing more than a ratio of two probabilities: the probability of an event happening divided by the probability that the event is not happening. For example, a probability of 10% corresponds to odds of 1/9, often described as 1 to 9. Conversely, 3 to 1 odds, say, correspond to a probability of $3/4 = 0.75$.[1]

We can now derive a simple version of Bayes' theorem in terms of odds, if we divide (31.1)–(31.2):

$$\frac{\Pr(D+ \mid T+)}{\Pr(D- \mid T+)} = \frac{\Pr(T+ \mid D+)}{\Pr(T+ \mid D-)} \times \frac{\Pr(D+)}{\Pr(D-)}. \tag{31.4}$$

We will refer to this equation as Bayes' theorem in *odds form*. Here $\Pr(D+)/\Pr(D-)$ are the *prior odds*, $\Pr(D+ \mid T+)/\Pr(D- \mid T+)$ are the *posterior odds* and $\Pr(T+ \mid D+)/\Pr(T+ \mid D-)$ is the so-called *likelihood ratio* for a positive test result, which we can easily identify as the sensitivity divided by 1 minus the specificity. Bayesian updating is thus just one simple mathematical operation: Multiply the prior odds with the likelihood ratio to obtain the posterior odds.

Example 31.1. (Continued)
The prior odds are 1/99 and the likelihood ratio (for a positive test result) is $0.9/0.1 = 9$. The posterior odds are therefore $9 \cdot 1/99 = 1/11 \approx 9.1\%$. So the prior odds of 1 to 99 change to posterior odds of 1 to 11 in the light of a positive test result. If the test result was negative, then the prior odds need to be multiplied with the likelihood ratio for a negative test result, which is $\Pr(T- \mid D+)/\Pr(T- \mid D-) = 0.1/0.9 = 1/9$. (Note that the likelihood ratio for a negative test result is 1 minus the sensitivity divided by the specificity.) This leads to posterior odds of $1/9 \cdot 1/99 = 1/891$. We leave it up to the reader to check that these posterior odds correspond to the positive and the negative predictive value, respectively, calculated earlier. Figure 31.1 illustrates Bayesian learning using odds in this example.

We now discuss an important formal aspect. Formula (31.1) is specified for a positive test result $T+$ and a diseased person $D+$ but is equally valid if we replace a positive test result $T+$ by a negative one, that is, $T-$, or $D+$ by $D-$, or both. In fact, we have already replaced $D+$ by $D-$ to write down (31.2).

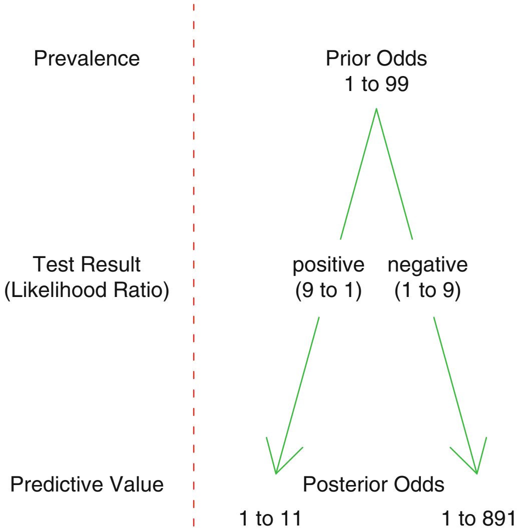
A more general description of Bayes' theorem is given by

$$\mathrm{p}(D = d \mid T = t) = \frac{\mathrm{p}(T = t \mid D = d) \times \mathrm{p}(D = d)}{\mathrm{p}(T = t)}, \tag{31.5}$$

where D and T are binary *random variables* which take the values d and t, respectively. In the diagnostic setting outlined above, d and t can be either "+"

[1] Odds $\omega = \pi/(1 - \pi)$ can be back-transformed to probabilities π using $\pi = \omega/(1 + \omega)$.

Fig. 31.1 Schematic representation of Bayesian learning in a diagnostic test example



or "–". Note that we have switched notation from Pr(.) to p(.) to emphasize that (31.5) relates to general *probability functions* of the random variables D and T, and not only to probabilities of the events $D+$ and $T+$, say.

An even more compact version of (31.5) is

$$\mathrm{p}(d \mid t) = \frac{\mathrm{p}(t \mid d) \times \mathrm{p}(d)}{\mathrm{p}(t)}. \tag{31.6}$$

Note that this equation also holds if the random variables D or T can take more than two possible values. The formula is also correct if it involves continuous random variables, in which case $\mathrm{p}(\cdot)$ denotes a density function.[2]

In reality, information on prevalence is typically estimated from a prevalence study while sensitivity and specificity are derived from a diagnostic study. However, the uncertainty associated with these estimates has been ignored in the above calculations. In the following, we will describe the Bayesian approach to quantify the uncertainty associated with these estimates. This can subsequently be used to assess the uncertainty of the positive and negative predictive values.

31.3 Bayesian Parameter Estimation

Conceptually, the Bayesian approach to parameter estimation treats all unknown quantities as random variables with appropriate prior distributions. Some knowledge of important elementary probability distributions is therefore required. Appendix B summarizes properties of the distributions used in this chapter.

[2] Probability statements for continuous random variables X can be obtained through integration of the density function, for example, $\Pr(a \leq X \leq b) = \int_a^b \mathrm{p}(x)dx$.

A *prior distribution* $\mathrm{p}(\theta)$ represents our knowledge about an unknown parameter θ, which we would like to update in the light of observed data x, whose probability of occurrence depends on θ. For example, x might be the results from an epidemiological study. The conditional probability function $\mathrm{p}(x \mid \theta)$ of x given θ is called the *likelihood function*. Combined with the prior distribution $\mathrm{p}(\theta)$, we can calculate the *posterior distribution* $\mathrm{p}(\theta \mid x)$ using Bayes' theorem:

$$\mathrm{p}(\theta \mid x) = \frac{\mathrm{p}(x \mid \theta) \times \mathrm{p}(\theta)}{\mathrm{p}(x)}. \tag{31.7}$$

This formula is of course just Eq. 31.6 with d replaced by θ and t replaced by x.

Note that the denominator $\mathrm{p}(x)$ does not depend on θ, its particular value is therefore not of primary interest. To compute the posterior distribution $\mathrm{p}(\theta \mid x)$ (up to a multiplicative constant), a simplified version of Bayes' theorem

$$\mathrm{p}(\theta \mid x) \propto \mathrm{p}(x \mid \theta) \times \mathrm{p}(\theta)$$

is sufficient.[3] In words, this formula says that the posterior distribution is proportional to the product of the likelihood function $\mathrm{p}(x \mid \theta)$ and the prior distribution $\mathrm{p}(\theta)$. Note that $\mathrm{p}(x \mid \theta)$, originally the probability or density function of the data x given the (unknown) parameter θ, is used as a function of θ for fixed x. It is convenient to write $\mathrm{L}_x(\theta)$ for $\mathrm{p}(x \mid \theta)$ to emphasize this fact:

$$\mathrm{p}(\theta \mid x) \propto \mathrm{L}_x(\theta) \times \mathrm{p}(\theta). \tag{31.8}$$

Note also that we need to know $\mathrm{L}_x(\theta)$ and $\mathrm{p}(\theta)$ only "up to scale," that is, we can ignore any multiplicative factors which do not depend on θ. This will often make the computations simpler.

A likelihood approach to statistical inference, see, for example, Pawitan (2001), uses only the likelihood $\mathrm{L}_x(\theta)$ and calculates the *Maximum Likelihood estimate* (MLE) $\hat{\theta}_{ML}$ defined as that particular value of θ which maximizes $\mathrm{L}_x(\theta)$. The likelihood function can also be used to compute frequentist *confidence intervals* based on the likelihood ratio test statistic. Alternatively, a Wald confidence interval can be calculated based on the *standard error* $\mathrm{se}(\hat{\theta}_{ML})$ of $\hat{\theta}_{ML}$, an estimate of the standard deviation of the MLE in (fictive) repeated experiments under identical conditions. The standard error can be calculated based on the curvature of the logarithm of the likelihood function (the so-called *log likelihood*) at the MLE.

In contrast, Bayes' theorem formalizes the fundamental principle of Bayesian inference in that the prior assumptions $\mathrm{p}(\theta)$ about θ are updated using the likelihood $\mathrm{p}(x \mid \theta)$ to obtain the posterior distribution $\mathrm{p}(\theta \mid x)$. The posterior distribution provides all information about the quantity of interest, but usually, we want to summarize it using point and interval estimates. The *posterior mean*, the mean of the posterior distribution, is the most commonly used point estimate, alternatively the

[3] The mathematical symbol $\propto$ stands for "is proportional to."

posterior median and the *posterior mode* may also be used. For interval estimation, any interval $[\theta_l, \theta_u]$ with $\Pr(\theta_l \leq \theta \leq \theta_u \,|\, x) = 1 - \alpha$ serves as a $(1 - \alpha) \cdot 100\%$ *credible interval*. A common choice is $\alpha = 5\%$, in which case we obtain 95% credible intervals. Often, *equi-tailed* credible intervals are used, where the same amount $(\alpha/2)$ of probability mass is cut-off from the left and right tail of the posterior distribution, that is,

$$\Pr(\theta < \theta_l \,|\, x) = \Pr(\theta > \theta_u \,|\, x) = \alpha/2.$$

So-called *highest posterior density* (HPD) intervals are also commonplace. They have the defining feature that the posterior density at any value of θ inside the credible interval must be larger than anywhere outside the credible interval. It can be shown that HPD credible intervals have the smallest width among all $(1-\alpha)$ credible intervals. If the posterior distribution is symmetric, for example, normal, then posterior mean, mode, and median coincide and HPD credible intervals are also equi-tailed. For non-symmetric posterior distributions, these quantities typically differ.

31.3.1 Choice of Prior Distribution

Compared with a classical approach to inference, the prior distribution has to be chosen appropriately, which often causes concerns to practitioners. In particular, a Bayesian analysis is often feared to introduce more unrealistic assumptions than a standard frequentist analysis. However, this is viewed as a misconception by many authors who consider the possibility to specify a prior distribution as something useful (Spiegelhalter et al. 2004; Greenland 2006, 2007).

The prior distribution should reflect the knowledge about the parameter of interest (e.g., a relative risk parameter in an epidemiological study). Ideally, this prior distribution should be elicited from experts (Spiegelhalter et al. 2004; O'Hagan et al. 2006). In the absence of expert opinions, simple *informative* prior distributions (e.g., that the relative risk parameter is with prior probability 95% between 0.5 and 2) may still be a reasonable choice. A sensitivity analysis with different prior distributions will help to examine how the conclusions depend on the choice of prior.

However, there have been various attempts to specify non-informative or reference priors to lessen the influence of the prior distribution. Reference priors used in such an "objective Bayes" approach typically correspond to rather unrealistic prior beliefs. However, "non-informative" priors provide a reference posterior where the impact of the prior distribution on the posterior distribution is minimized. Quite interestingly, such reference analyses may have equally good or even better frequentist properties than truly frequentist procedures (Bayarri and Berger 2004).

The most commonly used reference prior is *Jeffreys' prior*, named after the British physicist Harold Jeffreys (1891–1989). He proposed a general device to derive a non-informative prior distribution for a given likelihood function. It is interesting that the resulting non-informative reference prior is not necessarily a uniform prior. In many cases it is *improper*, that is, it does not sum or integrate to unity. For example, if the parameter θ can be any non-negative integer (without an upper limit) and we assume the same prior probability for each possible value of θ, then, this constitutes an improper prior distribution. In contrast, a *proper* prior will be a proper distribution in the mathematical sense. A proper prior can be easily achieved in this example by fixing an upper limit, that is, setting the prior probability of all values above that upper limit to zero. Operationally, improper priors are not a problem for parameter estimation, but they do cause problems in model selection, as we will see in Sect. 31.4. We will see some examples of Jeffreys' prior in the following.

31.3.2 Bayesian Analysis of Prevalence Data

The prevalence π is defined as the proportion of people in a population that has a specific disease. A simple prevalence study selects a random sample of n individuals from that population and counts the number x of diseased individuals. If the number of people in the population is large, then a binomial model $X \mid \pi \sim Bin(n, \pi)$[4] is appropriate to describe the statistical variability occurring in such a study design, see Appendix B for properties of the binomial distribution. Note that the MLE of π is $\hat{\pi}_{ML} = x/n$.

It is commonplace to select a beta distribution as prior distribution for π, because the beta distribution can only take values within the unit interval, that is, within the range of possible values of π. So assume that $\pi \sim Be(\alpha, \beta)$ a priori with $\alpha, \beta > 0$. Properties of the beta distribution are listed in Appendix B. Multiplying the binomial likelihood

$$\mathrm{L}_x(\pi) \propto \pi^x (1-\pi)^{n-x}$$

with the beta prior density

$$\mathrm{p}(\pi) \propto \pi^{\alpha-1}(1-\pi)^{\beta-1},$$

one easily obtains the posterior density

$$\begin{aligned}\mathrm{p}(\pi \mid x) &\propto \mathrm{L}_x(\pi) \times \mathrm{p}(\pi)\\ &\propto \pi^{\alpha+x-1}(1-\pi)^{\beta+n-x-1},\end{aligned}$$

compare Eq. 31.8. This can easily be identified as yet another beta distribution with parameters $\alpha + x$ and $\beta + n - x$:

$$\pi \mid x \sim Be(\alpha + x, \beta + n - x). \tag{31.9}$$

[4] The mathematical symbol $\sim$ stands for "is distributed as."

Compared with the prior distribution $\pi \sim Be(\alpha, \beta)$, the number of successes x is added to the first parameter while the number of failures $n - x$ is added to the second parameter. Note that the beta distribution is called *conjugate* to the binomial likelihood since the posterior is also beta distributed. Both prior and posterior density function are displayed in Fig. 31.2 for a simple example.

It is convenient to think of the $Be(\alpha, \beta)$ prior distribution as that which would have arisen had we started with an improper $Be(0, 0)$ prior and then observed α successes in $\alpha + \beta$ trials. Thus, $n_0 = \alpha + \beta$ can be viewed as a *prior sample size* and $\alpha/(\alpha + \beta)$ is the *prior mean*. This interpretation of the prior parameters is useful in order to intuitively assess the weight attached to the prior distribution, as we will see soon. It also stresses an important feature of Bayesian inference, the consistent processing of sequentially arising data. Indeed, suppose new independent data x_2 from the same $Bin(n, \pi)$ distribution arrives, then the posterior distribution following the original observation (with x now renamed to x_1) becomes the prior for the next observation x_2:

$$\mathrm{p}(\pi \mid x_1, x_2) \propto \mathrm{p}(x_2 \mid \pi) \times \mathrm{p}(\pi \mid x_1).$$

Here, we have been able to replace $\mathrm{p}(x_2 \mid \pi, x_1)$ by $\mathrm{p}(x_2 \mid \pi)$ due to the conditional independence of x_1 and x_2, given π. Now, $\mathrm{p}(\pi \mid x_1)$ is of course proportional to $\mathrm{p}(x_1 \mid \pi) \times \mathrm{p}(\pi)$, so an alternative formula is

$$\begin{aligned} \mathrm{p}(\pi \mid x_1, x_2) &\propto \mathrm{p}(x_2 \mid \pi) \times \mathrm{p}(x_1 \mid \pi) \times \mathrm{p}(\pi) \\ &= \mathrm{p}(x_1, x_2 \mid \pi) \times \mathrm{p}(\pi). \end{aligned}$$

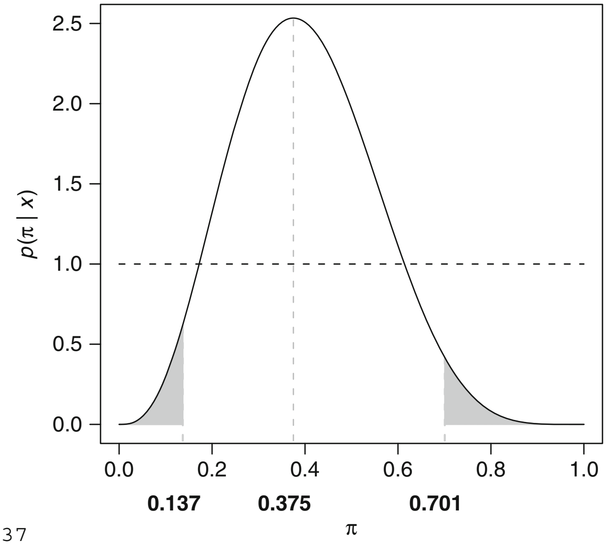


Fig. 31.2 A $Be(4, 6)$ posterior density $\mathrm{p}(\pi \mid x)$ (*solid line*) obtained from combining a $Be(1, 1)$ prior density (*dashed line*) with an observation $x = 3$ in a binomial experiment with $n = 8$ trials. The posterior mean is 0.4 and the posterior mode is 0.375. The equi-tailed 95% credible interval with limits $\theta_l = 0.137$ and $\theta_u = 0.701$ is also shown. The limits are calculated using the `R` function qbeta, that is, `qbeta(0.025,4,6) = 0.137`

In other words, $p(\pi \mid x_1, x_2)$ is the same whether or not the data are processed sequentially. Cornfield (1966, 1976) discusses this issue extensively in the context of clinical trials, see also Spiegelhalter et al. (2004, Sect. 4.3.2). Viewed from that perspective, every prior distribution is a posterior distribution based on the information available prior to the processing of the current data, and it makes sense to speak of a prior sample size. Bayesian inference is therefore sometimes described as Bayesian *learning*, which emphasizes the sequential nature inherent in the approach.

We now return to the posterior distribution in the binomial experiment. There are particularly simple explicit formulae for the mean and mode of a $Be(\alpha, \beta)$-distribution, see Appendix B for details. For example, the mean is simply $\alpha/(\alpha + \beta)$. Therefore, the posterior mean of $\pi \mid x \sim Be(\alpha + x, \beta + n - x)$ is

$$\frac{\alpha + x}{\alpha + \beta + n}.$$

Rewriting this as

$$\frac{\alpha + x}{\alpha + \beta + n} = \frac{\alpha + \beta}{\alpha + \beta + n} \cdot \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \cdot \frac{x}{n}$$

shows that the posterior mean is a weighted average of the prior mean $\alpha/(\alpha + \beta)$ and the MLE $\bar{x} = x/n$ with weights proportional to the prior sample size $n_0 = \alpha + \beta$ and the data sample size n, respectively. This further supports the interpretation of n_0 as a prior sample size. The *relative prior sample size* $n_0/(n_0+n)$ quantifies the weight of the prior mean in the posterior mean. Note that the relative prior sample size decreases with increasing data sample size n.

The case $\alpha = \beta = 1$ is of particular interest, as it corresponds to a uniform prior distribution on the interval $(0, 1)$, a natural "non-informative" choice. The prior sample size n_0 is now 2, one success and one failure. This is in fact exactly the prior used by Thomas Bayes (1702–1761) in his famous essay (Bayes 1763). The posterior mean is now $(x + 1)/(n + 2)$ and the posterior mode equals the MLE $\bar{x}$.

Incidentally, we note that Jeffreys' reference prior is not the uniform prior for π, but a beta distribution with both parameters α and β equal to 1/2, that is, $p(\pi) \propto \pi^{-0.5}(1 - \pi)^{-0.5}$ (compare Appendix B). This prior is proper and favors extreme values of π, that is, those which are close to either zero or one. From the above, we observe that Jeffreys' prior sample size n_0 is 1, half a success and half a failure.

Example 31.1. (Continued)
We now revisit the diagnostic test example discussed in Sect. 31.2 under the more realistic scenario that the disease prevalence $\pi = \Pr(D+)$ is not known, but only estimated from a prevalence study. For example, suppose there was $x = 1$ diseased individual in a random sample of size $n = 100$. Using a uniform $Be(1, 1)$ prior, the posterior distribution of π is Be(2,100), with posterior mean $1/51$ and posterior mode $1/100$.

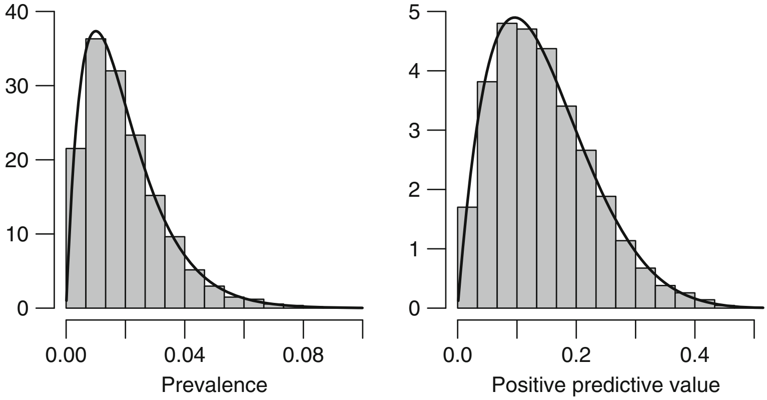


Fig. 31.3 *Left*: Posterior distribution of the prevalence. *Right*: Posterior distribution of the positive predictive value. Shown are histograms based on samples from these distributions. The *solid bold line* is the exact posterior density

It is tempting to replace the fixed prevalence $\Pr(D+) = 1/100$ in (31.1) with this $Be(2, 100)$ distribution to acknowledge the uncertainty involved in the prevalence estimation. The positive predictive value then follows a particular distribution, which can be computed analytically. However, it is much simpler to generate a random sample from the distribution of the positive predictive value using samples from the posterior distribution of π (Mossman and Berger 2001; Bayarri and Berger 2004). The following R-code illustrates this. Histograms of $n = 10{,}000$ samples from the prevalence and the associated positive predictive value are shown in Fig. 31.3.

```
> nsamples = 10000
> prev = rbeta(nsamples, 2, 100)
> sens = 0.9
> spec = 0.9
> ppv = sens * prev/(sens * prev + (1 - spec) * (1 - prev))
```

It is interesting that there is quite large uncertainty about the positive predictive value with 95% equi-tailed credible interval [0.02,0.34], which can be calculated from the corresponding quantiles of the sample. The 95% HPD credible interval is [0.009,0.31], so shifted to the left and slightly narrower, as expected. Note that the posterior mean is 0.145 (14.5%) and the posterior mode is 0.096 (9.6%). Both are larger than the positive predictive value 8.3% obtained for a fixed prevalence $\Pr(D+) = 0.01$ (see Sect. 31.2).

Mossman and Berger (2001) have considered a more general scenario where the characteristics of the diagnostic test are also not known exactly but based on estimates from a diagnostic study. Then, sensitivity `sens` and specificity `spec` in the above R-code needs to be replaced by corresponding samples from suitable beta (posterior) distributions, and the positive predictive value will have even more variation.

31.3.3 Bayesian Analysis of Incidence Rate Data

Incidence rate data typically consist of the number of cases x observed over the total person-time e. Alternatively, e may represent the expected number of cases under a specific assumption. A common approach, see, for example, Rothman (2002), is to assume that $X \sim Po(e\lambda)$, that is, the number of cases X is Poisson distributed with mean $e\lambda$ where e is a known constant and $\lambda > 0$ is an unknown parameter. If e is person-time, then λ represents the unknown incidence rate and if e is the number of expected cases, then λ is the unknown *rate ratio*, also called the *relative risk*.

It is commonplace to select a gamma distribution $Ga(\alpha, \beta)$ as prior distribution for λ, because it is conjugate to the Poisson likelihood, see Appendix B for details on the gamma distribution. The likelihood function of a Poisson observation with mean $e\lambda$ is

$$\mathrm{L}_x(\lambda) \propto \lambda^x \exp(-e\lambda).$$

It is easy to show that the MLE of λ is $\hat{\lambda}_{ML} = x/e$. Combining $\mathrm{L}_x(\lambda)$ with the density

$$\mathrm{p}(\lambda) \propto \lambda^{\alpha-1} \exp(-\beta\lambda)$$

of the gamma $Ga(\alpha, \beta)$ prior distribution, one obtains the posterior distribution of λ:

$$\begin{aligned}\mathrm{p}(\lambda \,|\, x) &\propto \mathrm{L}_x(\lambda) \times \mathrm{p}(\lambda) \\ &\propto \lambda^{\alpha+x-1} \exp(-(\beta + e)\lambda).\end{aligned}$$

This can be identified as another gamma distribution with parameters $\alpha + x$ and $\beta + e$:

$$\lambda \,|\, x \sim Ga(\alpha + x, \beta + e). \tag{31.10}$$

Compared with the prior distribution $Ga(\alpha, \beta)$, the number of observed counts x are added to the first parameter and the number of expected counts e are added to the second parameter.

The mean of a $Ga(\alpha, \beta)$ distribution is α/β, so the posterior mean is

$$\frac{\alpha + x}{\beta + e} = \frac{\beta}{\beta + e} \cdot \frac{\alpha}{\beta} + \frac{e}{\beta + e} \cdot \frac{x}{e}.$$

This equation illustrates that the posterior mean can be written as a weighted average of the prior mean α/β and the Maximum Likelihood estimate x/e with weights proportional to β and e, respectively. Hence, for Poisson data, there is a similar decomposition of the posterior mean as in the binomial case, see Sect. 31.3.2. Note that now β can be interpreted as prior sample size n_0 while e represents the data sample size.

Example 31.2. Breast cancer after fluroscopic examinations of the chest
For illustration, consider an example taken from Boice and Monson (1977), see also Greenland and Rothman (2008). A total of $x = 41$ breast cancer cases have been reported in a cohort of women treated for tuberculosis with x-ray fluoroscopy. Only $e = 23.3$ cases

were expected based on age-specific rates among women in Connecticut. We are interested in the posterior distribution of the rate ratio λ.

As prior distribution for the rate ratio λ, we may assume a gamma distribution with $\alpha = \beta$, and hence a prior mean of 1.0, that is, a prior expectation of a breast cancer rate after exposure to x-ray fluoroscopy equal to the overall rate in Connecticut. With a specific choice of α, we specify a range of plausible values around 1.0 which we consider believable a priori. For example, for $\alpha{=}\beta{=}8.78$, we believe that λ is in the range $[0.5, 2]$ with approximately 95% probability. Using this prior and Eq. 31.10, we obtain the posterior distribution $\lambda \mid x \sim Ga(8.78{+}41, 8.78{+}23.3) = Ga(49.78, 32.08)$. Note that the relative prior sample size is $8.78/32.08 \approx 27\%$, so the selected prior does have some weight in the posterior distribution.

The posterior mean of the relative risk λ is $49.78/32.08 = 1.55$. The equi-tailed 95% posterior credible interval for λ is $[1.15, 2.01]$. Thus, there is some evidence of an excess of breast cancers among x-rayed women relative to the reference group, but with quite large uncertainty about the actual size of the effect.

The above prior may be criticized for placing too much prior weight on relative risk values below unity with $\Pr(\lambda < 1) = 0.54$ and $\Pr(\lambda < 0.5) = 0.04$, but only $\Pr(\lambda > 2) = 0.01$. As a possible remedy, one may still pick a gamma prior with probability mass of 95% in the interval $[0.5, 2]$; however, one might want achieve symmetry by choosing a gamma prior which fulfills $\Pr(\lambda < 0.5) = \Pr(\lambda > 2) = 0.025$. This leads to the prior parameters $\alpha = 8.50$ and $\beta = 7.50$ and the posterior distribution $\lambda \mid x \sim Ga(49.50, 30.80)$. The posterior mean of the relative risk λ is now 1.61. The equi-tailed 95% posterior credible interval for λ is $[1.19, 2.08]$. The new prior gives very similar results compared to the original one.

Jeffreys' prior for the Poisson likelihood is a gamma distribution with parameters $\alpha = 1/2$ and $\beta = 0$. Since $\beta = 0$, it is an improper distribution but the associated posterior will be proper as long as $e > 0$. In the above example, the posterior mean 1.78 under Jeffreys' prior is larger as well as the limits of the 95% credible interval, which are 1.28 and 2.36.

31.3.4 Bayesian Analysis of Case-Control Data

We now turn to the Bayesian analysis of counts in a 2×2 table, with particular focus on the analysis of case-control studies with dichotomous exposure. Let E denote exposure and D disease, so let $\pi_1 = \Pr(E{+} \mid D{+})$ denote the probability that a case was exposed and $\pi_0 = \Pr(E{+} \mid D{-})$ the corresponding probability for a control. Assuming independent cases and controls and suitable (independent) priors for π_1 and π_0, we can easily derive the corresponding posterior distributions of π_1 and π_0, which are still independent. We may now proceed to infer the posterior distribution of the *odds ratio* $[\pi_1/(1 - \pi_1)]/[\pi_0(1 - \pi_0)]$. Conceptually, this is a simple mathematical problem; however, analytical calculation of the posterior density can be quite tedious (Nurminen and Mutanen 1987), so we use a simple Monte Carlo approach instead. The method gives independent samples from the posterior distribution of the odds ratio, as illustrated in the following example.

Example 31.3. Childhood leukemia and residential magnetic fields
Consider case-control data from Savitz et al. (1988), as reported in Greenland (2008), investigating a possible association between residential magnetic fields and childhood leukemia. For simplicity, the exposure variable was dichotomized based on a threshold of 3 milligauss (mG) exposure. The data are shown in Table 31.1. The entries of the 2×2 table are denoted by x_1 and y_1 for the exposed and unexposed cases, respectively, and x_0 and y_0

Table 31.1 Case-control data on residential magnetic field exposure and childhood leukemia

		Exposure	
		Exposed	Unexposed
Disease status	Cases	$x_1 = 3$	$y_1 = 5$
	Controls	$x_0 = 33$	$y_0 = 193$

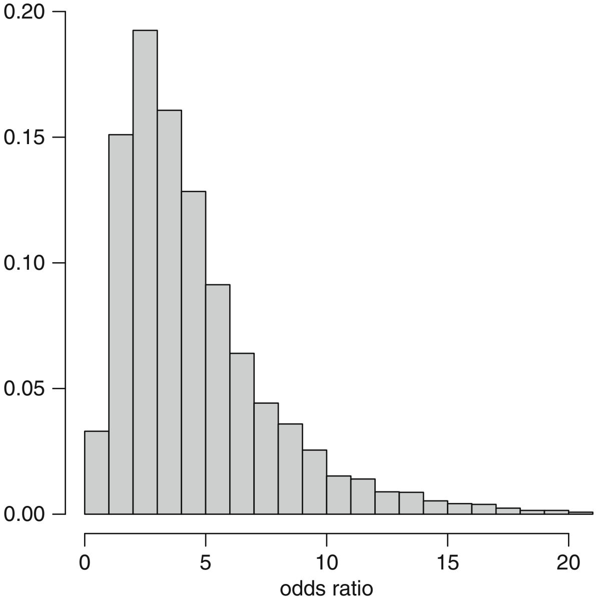


Fig. 31.4 A histogram based on 10,000 samples from the posterior odds ratio

for the corresponding controls. For simplicity, we continue to use the generic symbol x to denote all the available data for both cases and controls.

We assume independent uniform prior distributions for π_1 and π_0. It then follows that π_1 and π_0 are also a posteriori independent: $\pi_1 \,|\, x \sim Be(x_1 + 1 = 4, y_1 + 1 = 6)$ and $\pi_0 \,|\, x \sim Be(x_0 + 1 = 34, y_0 + 1 = 194)$, compare Sect. 31.3.2. In fact, Fig. 31.2 has shown the posterior of π_1. The following R-code illustrates a Monte Carlo approach to generate random samples from the posterior distribution of the odds ratio.

```
> nsamples = 10000
> pi1 = rbeta(nsamples, 4, 6)
> pi0 = rbeta(nsamples, 34, 194)
> or = (pi1/(1 - pi1))/(pi0/(1 - pi0))
```

Figure 31.4 gives a histogram of the posterior samples from the odds ratio `or`. The resulting posterior mean of the odds ratio is 4.7 with equi-tailed 95% credible interval [0.9,14.3]. Thus, there is large uncertainty about the odds ratio with values around unity not completely unrealistic.

The odds ratio $[\pi_1/(1-\pi_1)]/[\pi_0/(1-\pi_0)]$ is the odds $\pi_1/(1-\pi_1)$ to be exposed for a case divided by the odds $\pi_0/(1-\pi_0)$ to be exposed for a control, the so-called *exposure odds ratio*. Of more practical interest is the *disease odds ratio*, that is, the odds to be a case if exposed divided by the odds to be a case if not exposed. As Jerome Cornfield (1912–1979) has shown (Cornfield 1951) through yet another application of Bayes' theorem, the exposure odds ratio is in fact equal to the disease odds ratio. Cornfield's proof is quite simple: Bayes theorem (31.4) in odds form gives

$$\frac{\Pr(D+\,|\,E+)}{\Pr(D-\,|\,E+)} = \frac{\Pr(E+\,|\,D+)}{\Pr(E+\,|\,D-)} \times \frac{\Pr(D+)}{\Pr(D-)}.$$

and likewise

$$\frac{\Pr(D+\,|\,E-)}{\Pr(D-\,|\,E-)} = \frac{\Pr(E-\,|\,D+)}{\Pr(E-\,|\,D-)} \times \frac{\Pr(D+)}{\Pr(D-)}.$$

Dividing the first through the second equation gives after some rearrangement

$$\frac{\Pr(D+\,|\,E+)/\Pr(D-\,|\,E+)}{\Pr(D+\,|\,E-)/\Pr(D-\,|\,E-)} = \frac{\Pr(E+\,|\,D+)/\Pr(E-\,|\,D+)}{\Pr(E+\,|\,D-)/\Pr(E-\,|\,D-)}.$$

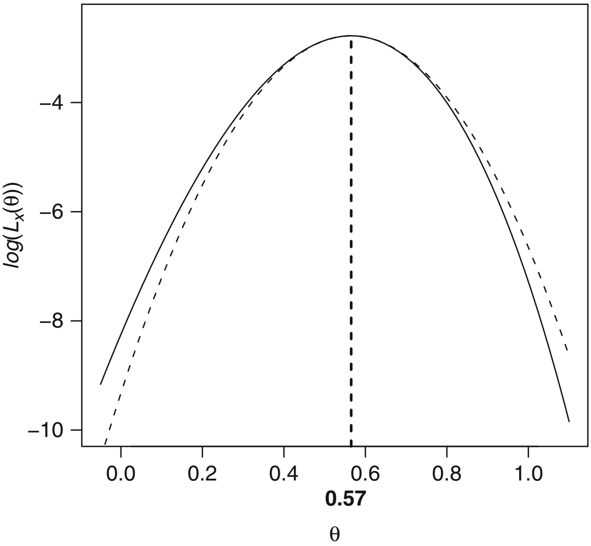
The left side of this equation is the disease odds ratio and the right side is the exposure odds ratio. For more details on statistical issues of the case-control study, see, for example, Breslow (1996) or chapter ▸Case-Control Studies of this handbook.

31.3.5 Approximate Bayesian Analysis

The shapes of many log likelihood functions $\log \mathrm{L}_x(\theta)$ are approximately quadratic, see, for example, Clayton and Hills (1993, Chap. 9). The log likelihood function of the normal distribution is exactly quadratic, and this fact can be used to apply techniques based on the normal distribution for approximate Bayesian inference (Greenland 2008). Methods based on approximate likelihoods are particularly important because the quadratic approximation becomes closer to the true likelihood as the sample size increases. Figure 31.5 illustrates this for the log likelihood of a Poisson observation $x = 41$ with mean $e \cdot \lambda$ where $e = 23.3$ (Example 31.2). Note that the log likelihood is shown not with respect to λ, but in terms of the log relative risk $\theta = \log(\lambda)$. The normal approximation is typically better if the parameter of interest is unrestricted, so it is better to approximate the log relative risk rather than the relative risk, which can take only positive values.

It is often appropriate to approximate a likelihood function of an unknown parameter θ by viewing the MLE as the actually observed (normal) data x. The associated standard error serves as (known) standard deviation $\sigma = \mathrm{se}(\hat{\theta}_{ML})$ of that normal distribution: $\hat{\theta}_{ML} \sim N(\theta, \sigma^2)$. The original likelihood function is hence replaced with its quadratic approximation, a likelihood function of one single normal observation x (the MLE) with known variance σ^2 (the squared standard

Fig. 31.5 Log likelihood function $\log \mathrm{L}_x(\theta)$ of a Poisson observation $x = 41$ with mean $e \cdot \exp(\theta) = 23.3 \cdot \exp(\theta)$ (*solid line*). Also shown is the MLE $\hat{\theta}_{ML} = 0.57$ and the quadratic approximation to the log likelihood (*dashed line*)



error). The unknown parameter θ is the mean of that normal distribution. Such an approach makes approximate Bayesian inference particularly simple, as we will see in the following.

So let X denote a sample from a normal $N(\theta, \sigma^2)$ distribution with mean θ and known variance σ^2. The corresponding likelihood function is

$$\mathrm{L}_x(\theta) \propto \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}.$$

Combined with a normal prior distribution for the unknown mean $\theta \sim N(\nu, \tau^2)$ with mean ν and variance τ^2, that is,

$$\mathrm{p}(\theta) \propto \exp\left\{-\frac{(\theta-\nu)^2}{2\tau^2}\right\}.$$

the posterior distribution is given by

$$\begin{aligned}\mathrm{p}(\theta \mid x) &\propto \mathrm{L}_x(\theta) \times \mathrm{p}(\theta)\\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{(x-\theta)^2}{\sigma^2} + \frac{(\theta-\nu)^2}{\tau^2}\right)\right\}.\end{aligned}$$

It can be shown that this is the density function of yet another normal distribution with variance $\tilde{\tau}^2 = 1/(1/\sigma^2 + 1/\tau^2)$ and mean $\tilde{\nu} = \tilde{\tau}^2(x/\sigma^2 + \nu/\tau^2)$:

$$\theta \mid x \sim N(\tilde{\nu}, \tilde{\tau}^2). \tag{31.11}$$

Table 31.2 A comparison of posterior characteristics for various prior distributions in the breast cancer study. Shown is the posterior mean $\hat{\lambda}$, the limits λ_l and λ_u of the equi-tailed 95% credible interval, and the tail probability $\Pr(\lambda < 1|x)$

Prior distribution	$\hat{\lambda}$	λ_l	λ_u	$\Pr(\lambda < 1\|x)$
$Ga(8.78, 8.78)$	1.55	1.15	2.01	0.00226
$Ga(8.5, 7.5)$	1.61	1.19	2.08	0.00116
$Ga(0.5, 0.0)$ (Jeffreys' prior)	1.78	1.28	2.36	0.00042
$LN(0, 0.125)$ (approximate)	1.62	1.21	2.12	0.00047
$LN(0, 0.125)$ (exact)	1.60	1.17	2.10	0.00170
$LN(0, \infty)$ (approximate)	1.78	1.30	2.39	0.00015

As for binomial samples, the posterior mean is a weighted mean of the prior mean ν and the data x with weights proportional to $1/\tau^2$ and $1/\sigma^2$, respectively. The relative prior sample size is thus $\tilde{\tau}^2/\tau^2$.

Example 31.2. (Continued)

It is well-known that the MLE $\hat{\theta}_{ML} = \log(x/e)$ of the log relative risk $\theta = \log(\lambda)$ is approximately normally distributed with mean equal to the true log relative risk θ and standard error $\sigma = 1/\sqrt{x}$ (Clayton and Hills 1993, Chap. 9). Using the data on breast cancer incidence after fluroscopic examinations of the chest from Sect. 31.3.3 where $x = 41$ and $e = 23.3$, the MLE of θ is hence 0.57 with standard error 0.16.

The MLE $\hat{\theta}_{ML}$ serves now as a summary of the information in the data to update our prior beliefs about θ. As prior distribution for θ, we select a mean-zero normal distribution such that the relative risk $\lambda = \exp(\theta)$ is between 0.5 and 2 with 95% probability. The corresponding normal distribution has variance $\tau^2 = (\log(2)/1.96)^2 \approx 1/8$. Note that a normal distribution for the log relative risk corresponds to a so-called *log-normal distribution* for the relative risk, where explicit formulae for the mean and mode are available (see Appendix B).

Using Eq. 31.11, the posterior variance is $\tilde{\tau}^2 = 1/(x+8) \approx 0.02$ and the posterior mean is $\tilde{\nu} = \tilde{\tau}^2(\hat{\theta}_{ML}/\sigma^2) = 0.47$. This corresponds to a posterior mean of $\exp(\tilde{\nu} + \tau^2/2) = 1.62$ for the relative risk λ (see the formula for the mean of a log-normal distribution in Appendix B). The associated 95% equi-tailed credible interval for the relative risk is $[1.21, 2.12]$. Note that the relative prior sample size is $\tilde{\tau}^2/\tau^2 \approx 0.16$, that is, 16%.

If we combine the exact Poisson likelihood with a normal prior for the log relative risk parameter θ, then the posterior distribution is no longer analytically tractable. However, posterior characteristics can be computed using numerical techniques. One obtains the posterior mean 1.60 and the 95% credible interval [1.17,2.10] for λ. These results are very similar to those based on the approximate analysis.

If we let the prior variance τ^2 of a normal prior $N(0, \tau^2)$ for the log relative risk parameter θ go to infinity, we obtain a "locally uniform" or flat prior, $p(\theta) \propto 1$, which is sometimes described as "non-informative." In this case, the posterior Eq. 31.11 simplifies to

$$\theta \mid x \sim N(x, \sigma^2). \tag{31.12}$$

Therefore, the point estimate of θ is simply the MLE $\hat{\theta}_{ML}$, and the limits of the equi-tailed 95% credible interval are numerically identical to the limits of the standard 95% Wald confidence interval:

$$\hat{\theta}_{ML} \pm 1.96 \cdot \sigma.$$

Results based on this approximate analysis with a flat prior for the log relative risk parameter θ are similar to the ones based on the reference prior for the Poisson mean λ, see Table 31.2. A standard frequentist analysis can thus be regarded as a Bayesian analysis using a reference prior. This connection between frequentist and Bayesian parameter estimates can be established in many other situations, at least approximately. However, viewed from a Bayesian perspective, the frequentist approach uses a rather unrealistic prior which gives large weight to unrealistically extreme values of relative risk.

31.3.6 Bayesian Tail Probabilities

In classical hypothesis testing, a commonly encountered procedure is the so-called *one-sided hypothesis test* (see, e.g., Cox (2005)) where the evidence against a null hypothesis $H_0 : \theta \leq \theta_0$ is quantified using a p-value:

$$p\text{-value } = \Pr(T(X) \geq T(x) \,|\, \theta = \theta_0),$$

here $T(X)$ is a suitable summary of the data X, for example, the mean. The p-value obtained from such a one-sided hypothesis test has sometimes a Bayesian interpretation as the posterior probability of H_0:

$$\Pr(H_0 \,|\, x) = \Pr(\theta < \theta_0 \,|\, x).$$

For illustration, consider a simple scenario with $n = 1$ observation $T(X) = X$ from a normal distribution with unknown mean θ and known variance σ^2. Under the assumption of a reference prior $\mathrm{p}(\theta) \propto 1$, the posterior distribution is $\theta \,|\, x \sim N(x, \sigma^2)$, see (31.12). Therefore,

$$\Pr(H_0 \,|\, x) = \Pr(\theta < \theta_0 \,|\, x) = \Phi((\theta_0 - x)/\sigma),$$

where Φ denotes the cumulative distribution function of the standard normal distribution. On the other hand, the p-value against H_0 is

$$p\text{-value } = \Pr(X \geq x \,|\, \theta = \theta_0) = 1 - \Phi((x - \theta_0)/\sigma) = \Phi((\theta_0 - x)/\sigma),$$

so is numerically equal to the posterior probability $\Pr(H_0 \,|\, x)$.

Of course, posterior probabilities can be calculated also for other prior distributions, in which case the analogy between posterior probabilities and p-values will usually be lost.

Example 31.2. (Continued)
Table 31.2 lists the posterior probability $\Pr(\lambda < 1 \,|\, x)$ for different prior assumptions on the relative risk parameter λ. It can be seen that there is some variation of these tail probabilities depending on the prior distribution and on the usage of an exact or an approximate approach, respectively. The frequentist p-value based on the Poisson distribution is $\Pr(X \geq 41 \,|\, \lambda = 1, e = 23.3) = 0.00057$, so within the range of the reported tail probabilities.

Note that the posterior probability $\Pr(\lambda < 1 \,|\, x) = 0.00047$ using the approximate approach is somewhat different from the corresponding one calculated with the exact likelihood, which is 0.00170. The reason for this discrepancy is the approximation of the Poisson log likelihood through a quadratic function, which corresponds to the approximate normal distribution of the log relative risk. Figure 31.5 shows that the quadratic approximation is good around the MLE, but not so good for small values of θ, with larger values of the log likelihood than its quadratic approximation. This explains the difference between the approximate and the exact results.

In the following, we show that the relationship between p-values and posterior tail probability may also hold (approximately) in quite unexpected circumstances. However, it is important to emphasize that the apparent analogy between p-values and posterior tail probabilities holds only in special cases and does not extend to the commonly used *two-sided* hypothesis test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, as we will see later.

31.3.6.1 A Tail Probability for Case-Control Data

In 1877, the medical doctor Carl von Liebermeister (1833–1901) proposed a Bayesian approach for the analysis of counts in a 2 × 2 table (Liebermeister 1877). Carl von Liebermeister was at that time professor in the Medical Faculty at the University of Tübingen in southern Germany. A Bayesian approach was selected by Liebermeister, as it was the inferential method of its time, following the tradition of Thomas Bayes and Pierre-Simon Laplace.

In the following, we will adopt the notation from Sect. 31.3.4 on the Bayesian analysis of case-control data, with π_1 and π_0 denoting the probability that a case and a control was exposed, respectively. Liebermeister had the ingenious idea to consider the posterior probability

$$\Pr(\pi_1 \leq \pi_0 \,|\, x) \tag{31.13}$$

in order to assess if there is evidence for a "significant" difference between cases and controls with respect to the underlying exposure risk. Liebermeister selected independent uniform priors for the unknown probabilities π_1 and π_0, directly following the approach by Thomas Bayes. Note that in modern epidemiological terminology, Eq. 31.13 is the posterior probability that the *relative risk* π_1/π_0 is smaller or equal to one. Furthermore, this probability is identical to the posterior probability that the odds ratio $\pi_1(1-\pi_0)/(\pi_0(1-\pi_1))$ is smaller or equal to one, because $\pi_1/\pi_0 \leq 1$ if and only if $\pi_1(1-\pi_0)/(\pi_0(1-\pi_1)) \leq 1$. Analytical computation of Eq. 31.13 is far from trivial, as reviewed in Seneta (1994). Quite interestingly, it turns out that Eq. 31.13 is the p-value of Fisher's one-sided exact test when testing the null hypothesis $\pi_1 \leq \pi_0$ against the one-sided alternative $\pi_1 > \pi_0$, if the diagonal entries x_1 and y_0 of the 2 × 2 table (here we adapt the notation from Table 31.1) are both increased by one count. Note that addition of 1 on the diagonal increases the empirical odds ratio and hence decreases the p-value of the above test.

The close connection to Fisher's test, which was developed more than 50 years later (Fisher 1934), has led Seneta (1994) to call the Liebermeister approach a "Bayesian test procedure." Seneta and Phipps (2001) studied frequentist properties of Eq. 31.13, viewed as a classical p-value. They showed that it has better average frequentist properties than the p-value obtained from Fisher's original test.

Altham (1969) has derived formulae for Eq. 31.13 in the more general setting of two independent beta distributions for π_0 and π_1 with arbitrary parameters. Nurminen and Mutanen (1987) have further generalized these results and have provided formulae for the whole posterior distribution of the risk difference, the risk ratio and the odds ratio. An interesting review of the Bayesian analysis of the 2 × 2 table can be found in Howard (1998). Note that all these authors have apparently been unaware of the original work by Liebermeister.

An alternative approximate approach for Bayesian inference in the 2 × 2 table, sometimes called *semi-Bayes*, has also been suggested. The basic idea is to re-parameterize the model in terms of a parameter of interest (e.g., the log odds ratio) and a so-called nuisance parameter (e.g., the log odds in the control group). A posterior distribution is then derived for the parameter of interest, assuming a suitable prior distribution. It is well known that the likelihood function for the log odds ratio ψ is approximately normal with mean $\log(x_1 \cdot y_0)/(x_0 \cdot y_1)$ and variance $(1/x_1 + 1/y_1 + 1/x_0 + 1/y_0)$, see Clayton and Hills (1993, Chap. 17). Adopting a flat (improper) prior for ψ, the posterior distribution is therefore also approximately normal with that mean and variance, which allows for the computation of (approximate) Bayesian p-values based on the normal distribution function. Proper normal priors can be easily incorporated in this approach using the techniques described in Sect. 31.3.5.

Example 31.3. (Continued)
Consider again the case-control example described in Sect. 31.3.4. The p-value from Fisher's one-sided test is 0.108, whereas Liebermeister's probability Eq. 31.13, calculated as the p-value of Fisher's test applied to Table 31.1 with diagonal entries increased to 4 and 194, respectively, is 0.036. Using the approximate approach with a flat improper reference prior for ψ, the posterior probability that the log odds ratio is equal to or smaller than zero (and hence the odds ratio is equal to or smaller than one) turns out to be 0.048, so quite similar. Greenland (2008) suggests an informative mean-zero normal prior distribution for ψ with variance 1/2. This distribution implies that the prior probability for an odds ratio between 1/4 and 4 is (approximately) 95%. Then, the posterior probability that the odds ratio is equal to or smaller than one is 0.127, so larger than before.

31.4 Prior Criticism and Model Choice

Various statistical researchers have emphasized the importance of modeling and reporting uncertainty in terms of *observables*, as opposed to inference about (unobservable) parameters. However, the latter, more traditional approach to inference can be seen as a limiting form of *predictive* inference about observables (Bernardo and Smith 1994). Parametric inference can therefore be seen as an intermediate structural step in the predictive process.

A predictive model for observables, for example, future outcomes of a clinical trial, can be constructed easily within the Bayesian framework. As we will see in this section, the prior predictive distribution plays also a key role in prior criticism and Bayesian model choice.

31.4.1 Bayesian Prediction

Suppose we want to predict future data x^{new}, say, which is assumed to arise from the same likelihood function as the original data x. Bayesian prediction is based on the simple identity

$$\mathrm{p}(x^{\text{new}} \mid x) = \int \mathrm{p}(x^{\text{new}} \mid \theta) \times \mathrm{p}(\theta \mid x)\, d\theta, \tag{31.14}$$

so the *predictive distribution* of x^{new} given x is the integral of the likelihood function of x^{new} times the posterior distribution $\mathrm{p}(\theta \mid x)$ with respect to θ.

For example, consider the binomial model with unknown success probability π as described in detail in Sect. 31.3.2. Suppose we want to predict a future observation $X^{\text{new}} \sim Bin(1, \pi)$. It is easy to show that $X^{\text{new}} \mid x$ has a Bernoulli distribution with success probability equal to the mean of $\pi \mid x$.

The predictive distribution (31.14) is sometimes called *posterior predictive distribution* since it is conditional on the observed data x. In contrast, the *prior predictive distribution*

$$\mathrm{p}(x) = \int \mathrm{p}(x \mid \theta) \times \mathrm{p}(\theta)\, d\theta \tag{31.15}$$

is derived from the likelihood and the prior distribution alone. The prior predictive distribution plays a key role in Bayesian model criticism and model selection, as we will see in the following section.

Note that calculation of the prior predictive distribution requires that $\mathrm{p}(\theta)$ is proper; otherwise, $\mathrm{p}(x)$ would be undefined. Note also that $\mathrm{p}(x)$ is the denominator in Bayes' theorem (31.7). Therefore,

$$\mathrm{p}(x) = \frac{\mathrm{p}(x \mid \theta) \times \mathrm{p}(\theta)}{\mathrm{p}(\theta \mid x)}, \tag{31.16}$$

which holds for any value of θ. This formula is very useful if both prior and posterior are available in closed form, in which case the integration in definition (31.15) can be avoided. However, it is necessary to include all normalizing constants in $\mathrm{p}(x \mid \theta)$, $\mathrm{p}(\theta)$, and $\mathrm{p}(\theta \mid x)$, which makes the calculations slightly more tedious.

31.4.2 Prior Criticism

Box (1980) has suggested an approach to compare priors with subsequent data. The method is based on a p-value obtained from the prior predictive distribution and the actually observed datum. Small p-values indicate a *prior-data conflict*, that is, incompatibility of prior assumptions and the actual observations.

Box's p-value is defined as the probability of obtaining a result with prior predictive ordinate $\mathrm{p}(X)$ equal to or lower than at the actual observation x:

$$\Pr(\mathrm{p}(X) \leq \mathrm{p}(x)).$$

If both data and prior are normal, $X \mid \theta \sim N(\theta, \sigma^2)$ and $\theta \sim N(\nu, \tau^2)$, then the prior predictive distribution is $X \sim N(\nu, \sigma^2 + \tau^2)$. It can be shown that Box's p-value is then the upper tail probability of a chi-squared distribution with 1 degree of freedom (a more common name for a $Ga(1/2, 1/2)$ distribution) evaluated at

$$v^2 = \frac{(x-\nu)^2}{\sigma^2 + \tau^2}.$$

Example 31.3. (Continued)
Let us revisit the case-control study from Sect. 31.3.4. The MLE of θ is $x = \log(3 \cdot 193/(5 \cdot 33)) \approx 1.255$ with standard error $\sigma = \mathrm{se}(\hat{\theta}_{ML}) = \sqrt{1/3 + 1/193 + 1/5 + 1/33} = 0.754$. Greenland's prior mean for the log odds ratio θ was $\nu = 0$ and the prior variance was $\tau^2 = 0.5$. We, hence, obtain

$$v^2 = \frac{1.255^2}{0.754^2 + 0.5} = 1.47$$

with an associated p-value of `1 - pgamma(1.47, 1/2, 1/2)`=0.22. Thus, by this check, the prior and the actually observed data appear to be fairly compatible, because Box's p-value is not remarkably small.

31.4.3 Bayesian Model Selection

Suppose we entertain two competing Bayesian models M_0 and M_1 and we are interested to know which one describes the data x better. Bayesian model choice is based on a variant of Eq. 31.4. Suppose we denote by $\Pr(M_0)$ and $\Pr(M_1)$ the prior probabilities of model M_0 and M_1, respectively, with $\Pr(M_0) + \Pr(M_1) = 1$. Then, the following fundamental equality holds:

$$\frac{\Pr(M_0 \mid x)}{\Pr(M_1 \mid x)} = \frac{\mathrm{p}(x \mid M_0)}{\mathrm{p}(x \mid M_1)} \times \frac{\Pr(M_0)}{\Pr(M_1)}. \tag{31.17}$$

Here, $\Pr(M_0)/\Pr(M_1)$ are the *prior odds*, $\Pr(M_0 \mid x)/\Pr(M_1 \mid x)$ are the *posterior odds*, and $\mathrm{p}(x \mid M_0)/\mathrm{p}(x \mid M_1)$ is the so-called *Bayes factor*, the ratio of the prior predictive distributions of the two models, both evaluated at the observed data x. The Bayes factor, which can be larger or smaller than one, summarizes the evidence of the data for the two models. If the Bayes factor is larger than one, then there is evidence for model M_0; otherwise, there is evidence for model M_1. Note that the Bayesian approach to model selection treats the two models M_0 and M_1 in a symmetric fashion, whereas classical hypothesis tests can only reject, but never accept the simpler model.

The term $\mathrm{p}(x \mid M)$ is also known as the *marginal likelihood*, to contrast it with the ordinary (conditional) likelihood $\mathrm{p}(x \mid \theta, M)$. The marginal likelihood can be calculated based on the prior predictive distribution (31.15).

Example 31.3. (Continued)
We revisit the approximate Bayesian analysis for case-control data and compare model M_0 with a fixed odds ratio of one with model M_1, where we use as before a $N(0, 0.5)$ prior for the log odds ratio ψ. This model comparison is the Bayesian version of the classical

two-sided hypothesis test for the null hypothesis that the odds ratio equals one. As before, we adopt an approximate Bayesian analysis assuming that the observed log odds ratio $\hat{\psi} = 1.26$ is normally distributed with known variance 0.57. The (marginal) likelihood in model M_0 is thus simply the density of a normal distribution with mean zero and variance 0.57, evaluated at $\hat{\psi} = 1.26$. This turns out to be 0.13. The prior predictive distribution in model M_1 is also normal with mean zero, but with variance $0.5 + 0.57 = 1.07$, so the marginal likelihood in model M_1 is 0.18. The Bayes factor of model M_0 relative to model M_1 is therefore $0.13/0.18 = 0.72$. Assuming 1 to 1 prior odds, the posterior odds for M_0 versus M_1 are therefore 0.72, and the corresponding posterior probability of model M_0 has decreased from 0.5 to $0.72/(1 + 0.72) = 0.42$ using the formula in Footnote 1 on page 1164.

It is somewhat surprising that the posterior probability has barely changed, despite a fairly small p-value obtained from Fisher's two-sided test ($p = 0.11$). The corresponding Wald test gives a similar result ($p = 0.096$). This illustrates that the correspondence between Bayesian model selection and p-values is typically lost for the standard two-sided hypothesis test (Berger and Sellke 1987). In particular, p-values cannot be interpreted as posterior probabilities of the null hypothesis.

In the following, we will study the two-sided hypothesis test $M_0 : \theta = 0$ versus $M_1 : \theta \neq 0$ in more detail, assuming that the MLE $\hat{\theta}_{ML}$ is normal distributed with unknown mean θ and known variance σ^2, equal to the squared standard error of $\hat{\theta}_{ML}$. This scenario reflects, at least approximately, many of the statistical procedures found in epidemiological journals.

We now have the possibility to calculate a *minimum Bayes factor* (MBF) (Edwards et al. 1963; Goodman 1999a,b), a lower bound on the evidence against the null hypothesis. The idea is to consider a whole family of prior distributions and to derive a lower bound for the Bayes factor in that family, the minimum Bayes factor. The approach can be taken to the limit by considering all possible prior distributions, in which case the minimum Bayes factor is a universal bound on the evidence of the data against the null hypothesis. Interestingly, the prior distribution $\mathrm{p}(\theta)$ in model M_1 with smallest Bayes factor is concentrated at the MLE $\hat{\theta}_{ML}$, that is, assumes $\theta = \hat{\theta}_{ML}$ a priori. If a z-value $z = \hat{\theta}_{ML}/\mathrm{se}(\hat{\theta}_{ML})$ has been calculated for this two-sided test, then the following formula can be used to calculate this universal minimum Bayes factor (Goodman 1999b):

$$\mathrm{MBF} = \exp\left(-\frac{z^2}{2}\right).$$

For example, if $z = 1.96$, where the two-sided p-value is 0.05, then $\mathrm{MBF} = \exp(-1.96^2/2) \approx 0.15$. If we assume 1 to 1 prior odds, then a universal lower bound on the posterior probability of the null hypothesis M_0 is therefore $0.15/(1+0.15) = 0.13$.

However, the above approach has been criticized since a prior distribution concentrated at the MLE is completely unrealistic since we do not know the MLE a priori. In addition, since the alternative hypothesis has all its prior density on one side of the null hypothesis, it is perhaps more appropriate to compare the outcome of this procedure with the outcome of a one-sided rather than a two-sided test, in which case MBF ≈ 0.26, so considerably larger.

Minimum Bayes factors can also be derived in more realistic scenarios. A particularly simple approach (Sellke et al. 2001) leads to the formula

$$\mathrm{MBF} = c \cdot p \log(p),$$

where $c = -\exp(1) \approx -2.72$ and p denotes the p-value from the two-sided hypothesis test (assumed to be smaller than $\exp(-1) \approx 0.37$). For example, for $p = 0.05$, we obtain MBF ≈ 0.41.

Example 31.3. (Continued)
In the case-control example, $z = 1.255/0.754 \approx 1.66$, and we obtain the minimum Bayes factor of $\exp(-1.66^2/2) \approx 0.25$, that is, 1 to 4, and a lower bound of 0.2 on the corresponding posterior probability of model M_0 (assuming 1 to 1 prior odds). Thus, it is impossible that the posterior probability of the null hypothesis is smaller than 0.2 if our prior probability was 0.5.

Of course, from the above, z-value a p-value can be easily calculated, which turns out to be $p = 0.096$. Using this p-value, the more realistic Sellke et al. (2001) approach gives a minimum Bayes factor of 0.61, which corresponds to a lower bound of 0.38 on the corresponding posterior probability. We conclude that for two-sided hypothesis tests, the evidence against the null hypothesis is by far not as strong as the p-value seems to suggest. This general finding is discussed extensively in the literature (Edwards et al. 1963; Berger and Sellke 1987; Goodman 1999a,b; Sellke et al. 2001; Goodman 2005).

31.5 Further Topics

We now discuss more advanced techniques of Bayesian inference: empirical Bayes approaches and Markov chain Monte Carlo methods.

31.5.1 Empirical Bayes Approaches

Empirical Bayes methods are a combination of the Bayesian approach with likelihood techniques. The general idea is to estimate parameters of the prior distribution $\mathrm{p}(\theta)$ from multiple experiments, rather than fixing them based on prior knowledge. Strictly speaking, this is not a fully Bayesian approach, but it can be shown that empirical Bayes estimates have attractive theoretical properties. Empirical Bayes techniques are often used in various applications. For a general discussion, see also Davison (2003, Sect. 11.5). Here, we sketch the idea in an epidemiological context discussing shrinkage estimates of age-standardized relative risks for use in disease mapping.

Suppose that for each region $i = 1, \ldots, n$ the observed number of cases x_i of a particular disease are available as well as the expected number e_i under the assumption of a constant disease risk. We now present a commonly used empirical Bayes procedure which is due to Clayton and Kaldor (1987).

Assume that $x_1, \ldots, x_n$ are independent realizations from $Po(e_i \lambda_i)$ distributions with known expected counts $e_i > 0$ and unknown region-specific parameters λ_i. A suitable prior for the λ_i's is a gamma distribution, $\lambda_i \sim Ga(\alpha, \beta)$, due to the conjugacy of the gamma distribution to the Poisson likelihood. The posterior of λ_i turns out to be

$$\lambda_i \mid x_i \sim Ga(\alpha + x_i, \beta + e_i) \tag{31.18}$$

with posterior mean $(\alpha + x_i)/(\beta + e_i)$, compare Sect. 31.3.3. If α and β are fixed in advance, the posterior of λ_i does not depend on the data x_j and e_j from the other regions $j \neq i$.

Empirical Bayes estimates of λ_i are based on (31.18), but the parameters α and β of the prior distribution are not fixed in advance but estimated based on all available data. This is done by maximizing the implied prior predictive distribution or marginal likelihood, which depends only on α and β. One obtains MLEs $\hat{\alpha}_{ML}$ and $\hat{\beta}_{ML}$ of α and β, which are plugged into Formula (31.18). The resulting posterior mean estimates

$$\frac{\hat{\alpha}_{ML} + x_i}{\hat{\beta}_{ML} + e_i} \tag{31.19}$$

are called *empirical Bayes estimates* of λ_i. They will always lie between the MLEs x_i/e_i and the estimated mean $\hat{\alpha}_{ML}/\hat{\beta}_{ML}$ of the gamma prior; thus, the MLEs are shrunk toward the common value $\hat{\alpha}_{ML}/\hat{\beta}_{ML}$. This phenomenon is called *shrinkage*.

Example 31.4. Lip cancer in Scotland
Consider data on the incidence of lip cancer in $n = 56$ regions of Scotland, as reported in Clayton and Kaldor (1987). Here we obtain $\hat{\alpha}_{ML} = 1.88$ and $\hat{\beta}_{ML} = 1.32$. Figure 31.6 displays the empirical Bayes estimates and the corresponding 95% equi-tailed credible intervals, ordered with respect to the MLEs. Figure 31.6 shows clearly that the MLEs x_i/e_i are shrunk to the prior mean, that is, the empirical Bayes estimates lie between these two extremes. A map of Scotland with the empirical Bayes estimates is shown in Fig. 31.7.

31.5.2 Markov Chain Monte Carlo

Application of ordinary Monte Carlo methods is difficult if the unknown parameter is of high dimension. However, *Markov chain Monte Carlo* (MCMC) methods will then be a useful alternative. The idea is to simulate a *Markov chain* $\theta^{(1)}, \ldots, \theta^{(m)}, \ldots$ in a specific way such that it converges to the posterior distribution $\mathrm{p}(\theta \mid x)$. After convergence, one obtains random samples from the target distribution, which can be used to estimate posterior characteristics as in ordinary Monte Carlo approaches. To ensure that the samples are taken from the target distribution, in practice, the first iterations, the so-called *burn-in*, are typically ignored. However, note that these samples will be dependent, an inherent feature of Markov chains.

The theory of MCMC is beyond the scope of this chapter, but we will illustrate the procedure in the context of disease mapping as discussed in Sect. 31.5.1. We

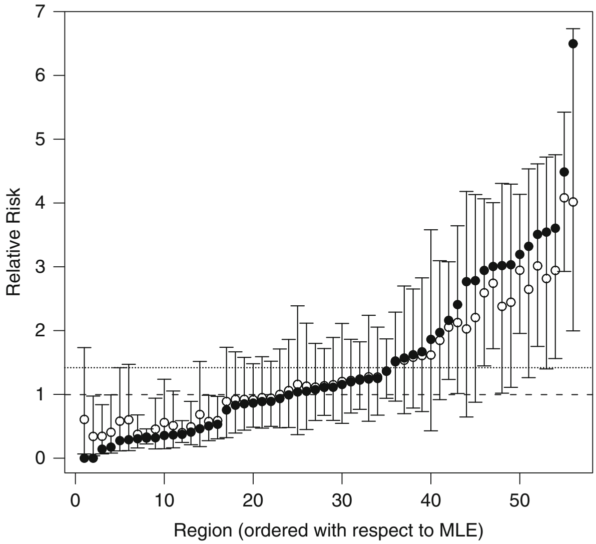


Fig. 31.6 Ninety-five percent equi-tailed credible intervals for Scottish lip cancer incidence rates λ_i $(i = 1, \ldots, 56)$, calculated with an empirical Bayes approach. The *dotted line* marks the MLE $\hat{\alpha}_{ML}/\hat{\beta}_{ML} = 1.42$ of the prior mean. *Open circles* denote the posterior mean estimates of λ_i. The regions are ordered with respect to their MLEs x_i/e_i, shown as *filled circles*

now specify a prior on the log relative risks $\theta_i = \log(\lambda_i)$ which takes into account spatial structure and thus allows for spatial dependence (Besag et al. 1991). More specifically, we use a *Gaussian Markov random field* (GMRF), most easily specified through the conditional distribution of θ_i given $\theta_{j \neq i}$, that is, the log relative risks in all other regions $j \neq i$. A common choice is to assume that

$$\theta_i \mid \theta_{j \neq i}, \tau^2 \sim N\left(\bar{\theta}_i, \frac{\tau^2}{n_i}\right), \tag{31.20}$$

here $\bar{\theta}_i = n_i^{-1} \sum_{j \sim i} \theta_j$ denotes the mean of the n_i spatially neighboring regions of region i and τ^2 is an unknown variance parameter. Some decision has to be made to connect the two islands shown in Fig. 31.7 to the rest of Scotland. Here, we assume that they are both adjacent to the nearest mainland region.

To simulate from the posterior distribution a specific MCMC approach, the *Gibbs sampler*, iteratively updates the unknown parameters $\theta_1, \ldots, \theta_n, \tau^2$. We omit details here but refer the interested reader to the relevant literature, for example, Rue and Held (2005).

Example 31.4. (Continued)
We now revisit the lip cancer incidence data in the $n = 56$ geographical regions of Scotland, allowing for spatial dependence between the relative risk parameters as described above. The following results are based on a Markov chain of length 100,000 where the first 10,000 samples were disregarded as burn-in. Figure 31.8 displays the corresponding posterior mean relative risks. Compared with the empirical Bayes estimates shown in Fig. 31.7, obtained from a model without spatial dependence, a spatially smoother picture can be observed.

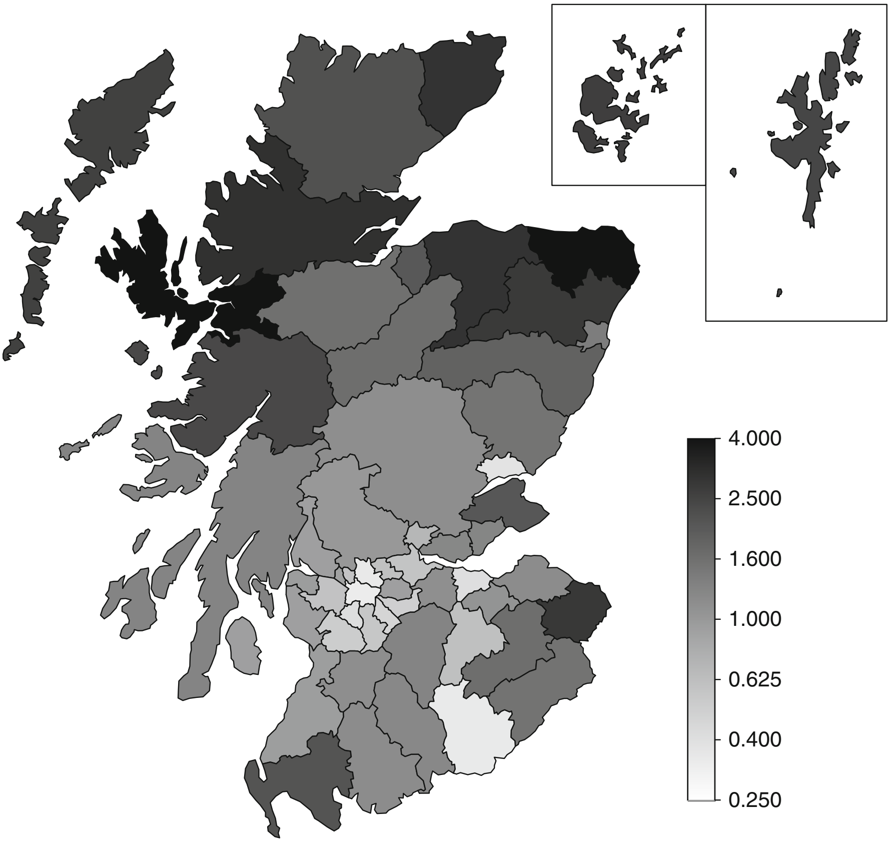


Fig. 31.7 Geographical distribution of the empirical Bayes estimates of the relative risk of lip cancer in Scotland

31.6 Conclusions

The Bayesian approach to statistical inference offers a coherent framework, in which both parameter estimation and model selection can be addressed. The key ingredient is the prior distribution, which reflects our knowledge about parameters or models before we integrate new data in our analysis. Bayesian statistics produces statements about the uncertainty of unknown quantities conditional on known data. This natural approach is in sharp contrast to frequentist procedures, which produce probability statements about hypothetical repetitions conditional on the unknown parameter and model.

The key to Bayesian statistics is the representation of prior beliefs through appropriate probability distributions. The key technique to update these prior beliefs in the light of new data is Bayes' theorem. Bayesian inference thus provides a

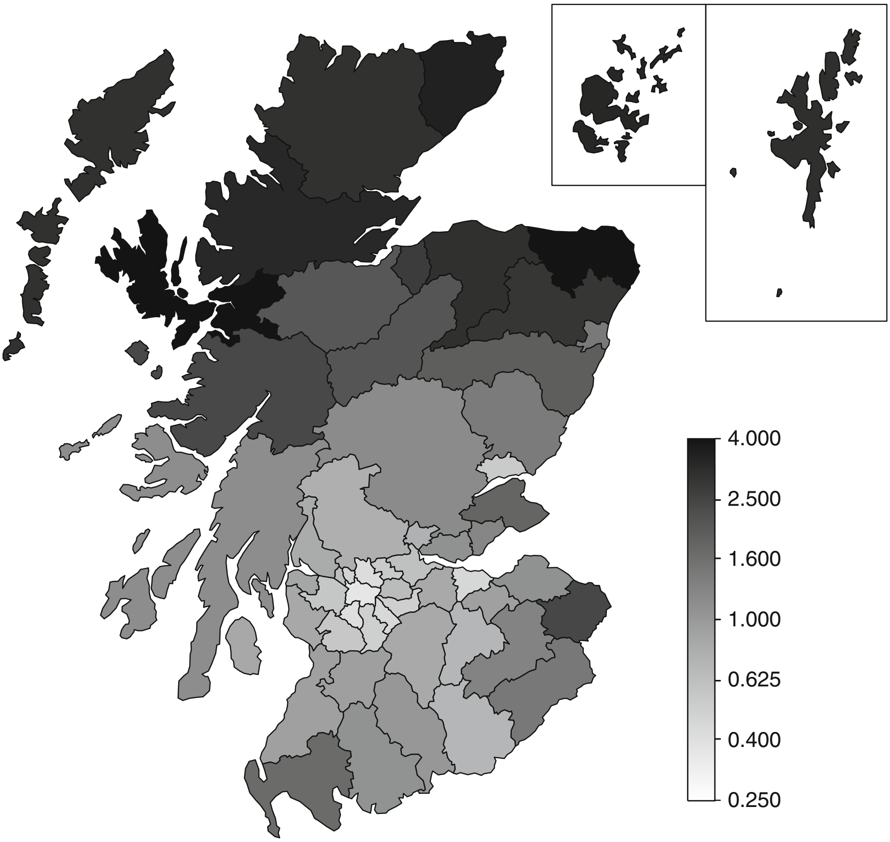


Fig. 31.8 Geographical distribution of the posterior mean relative risk estimates of lip cancer in Scotland, obtained through a GMRF approach

coherent way to update our knowledge in the light of new data. In the absence of conjugacy, the computation of the posterior distribution may require certain advanced numerical techniques such as Markov chain Monte Carlo.

I think it is important to emphasize relationships and differences between the frequentist and the Bayesian approach in order to appreciate what each of the different inference schools has to offer. A frequentist approach to parameter estimation based on the likelihood function alone can be regarded as a limited Bayesian form of inference in the absence of any prior knowledge. Such an approach typically leads to numerically similar results of both point and interval estimates. However, the possibility to specify a prior distribution is increasingly considered as something useful, avoiding implicit unrealistic assumptions of a frequentist analysis. Empirical Bayes approaches, which estimate a prior distribution from multiple experiments, are a compromise between the frequentist and Bayesian approach.

However, the frequentist and the Bayesian approach can lead to very different answers when it comes to hypothesis testing. A striking example is the two-sided hypothesis test, where the evidence against the null hypothesis, quantified by the Bayes factor, is by far not as strong as the p-value might suggest.

Appendix A. Rules of Probability

In this appendix, we summarize basic rules from probability theory. We also give a summary of important probability distributions.

A.1. Probabilities and Conditional Probabilities

Any experiment involving randomness can be modeled with probabilities. Probabilities are assigned to events such as "It will be raining tomorrow" or "I will suffer a heart attack in the next year." The *certain event* has probability one while the *impossible event* has probability zero. From a Bayesian perspective, probabilities are subjective in the sense that they quantify personal uncertainty that the event considered actually happens. Subjective probabilities can be elicited with a simple bet. If the actual realization of the event considered gives a return of 100 US dollar, say, and somebody is willing to bet up to but not more than p US dollars on that event happening, then his personal probability for the event is $p/100$.

Any event A has a disjoint, *complementary event* A^c such that $\Pr(A)+\Pr(A^c) = 1$. For example, if A is the event that "It will be raining tomorrow" then A^c is the event that "It will be not raining tomorrow." More generally, a series of events $A_1, A_2, \ldots, A_n$ is called a *partition* if the events are pairwise disjoint and if $\Pr(A_1) + \ldots + \Pr(A_n) = 1$.

Conditional probabilities $\Pr(A \mid B)$ are calculated to update the probability $\Pr(A)$ of a particular event under the additional information that a second event B has occurred. They can be calculated via

$$\Pr(A \mid B) = \frac{\Pr(A, B)}{\Pr(B)}, \tag{A.1}$$

where $\Pr(A, B)$ is the probability that both A and B occur. Rearranging this equation gives $\Pr(A, B) = \Pr(A \mid B) \Pr(B)$, but $\Pr(A, B) = \Pr(B \mid A) \Pr(A)$ must obviously also hold. Equating and rearranging these two formulas gives Bayes' theorem:

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \Pr(A)}{\Pr(B)}. \tag{A.2}$$

Conditional probabilities behave like ordinary probabilities if the conditional event is fixed, so $\Pr(A \mid B) + \Pr(A^c \mid B) = 1$. It then follows that

$$\Pr(B) = \Pr(B \mid A)\Pr(A) + \Pr(B \mid A^c)\Pr(A^c), \tag{A.3}$$

and more generally,

$$\begin{aligned}\Pr(B) = \Pr(B \mid A_1)\Pr(A_1) + \Pr(B \mid A_2)\Pr(A_2) + \ldots \\ \ldots + \Pr(B \mid A_n)\Pr(A_n)\end{aligned} \tag{A.4}$$

if $A_1, A_2, \ldots, A_n$ is a partition. This is called the *law of total probability*. Equation (A.3) and (A.4) may be useful to calculate the denominator in Eq. (A.2).

A.2. Probability Functions

We now switch notation and replace Pr with p and events A and B with possible realizations x and y of *random variables* X and Y to indicate that the rules described in Appendix A.1 hold for any event considered. The formulas also hold if continuous random variables are considered, in which case $\mathrm{p}(\cdot)$ is a *density function*. For example, Eq. (A.1) now reads

$$\mathrm{p}(x \mid y) = \frac{\mathrm{p}(x, y)}{\mathrm{p}(y)} \tag{A.5}$$

while Bayes' theorem (A.2) translates to

$$\mathrm{p}(x \mid y) = \frac{\mathrm{p}(y \mid x)\,\mathrm{p}(x)}{\mathrm{p}(y)}. \tag{A.6}$$

Similarly, the law of total probability (A.4) now reads

$$\mathrm{p}(y) = \int \mathrm{p}(y \mid x)\,\mathrm{p}(x) dx, \tag{A.7}$$

where the integral $\int dx$ with respect to x is to be understood as a sum over x if $\mathrm{p}(x)$ is the probability function of a discrete random variable X. Combining equations (A.5) and (A.7) shows that the variable x has to be integrated out of the joint density $\mathrm{p}(x, y)$ to obtain the marginal density $\mathrm{p}(y)$ of Y:

$$\mathrm{p}(y) = \int \mathrm{p}(x, y)\, dx. \tag{A.8}$$

Appendix B. Important Probability Distributions

The following table gives some elementary facts about the probability distributions used in this chapter. A random variable is denoted by X, and its probability or density function is denoted by $\mathrm{p}(x)$. For each distribution, the mean $\mathrm{E}(X)$, variance $\mathrm{var}(X)$, and mode $\mathrm{mod}(X)$ is listed, if appropriate.

In the first row, we list the name of the distribution, an abbreviation, and the core of the corresponding `R`-function (*e.g.* `_norm`). Depending on the first letter, represented by the placeholder "`_`," these functions can be conveniently used as follows:

`r` stands for *r*andom and generates independent random numbers from that distribution. For example, `rnorm(n, mean = 0, sd = 1)` generates n random numbers from the standard normal distribution.

`d` stands for *d*ensity and returns the probability and density function, respectively. For example, `dnorm(x)` gives the density of the standard normal distribution.

`p` stands for *p*robability and gives the so-called *distribution function* of X. For example, if X is standard normal, then `pnorm(0)` returns 0.5 while `pnorm(1.96)` is 0.975.

`q` stands for *q*uantile and gives the *quantile function*. For example, `qnorm(0.975)` is $1.959964 \approx 1.96$.

The first argument `arg` depends on the particular function used. It is either the number n of random variables generated, a value x in the domain of the random variable, or a probability p with $0 < p < 1$.

Binomial: $Bin(n, \pi)$	`_binom(arg, size =` n`, prob =` π`)`
$0 < \pi < 1, n \in \{1, \dots, n\}$	$x \in \{0, \dots, n\}$
$\mathrm{p}(x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}$	$\mathrm{L}_x(\pi) \propto \pi^x(1-\pi)^{n-x}$
$\mathrm{E}(X) = n\pi$	$\mathrm{var}(X) = n\pi(1-\pi)$
If $n = 1$ one obtains the *Bernoulli distribution*.	

Poisson: $Po(\lambda)$	`_pois(arg, lambda =` λ`)`
$\lambda > 0$	$x \in \{0, 1, \dots\}$
$\mathrm{p}(x) = \frac{\lambda^x}{x!}\exp(-\lambda)$	$\mathrm{L}_x(\lambda) \propto \lambda^x \exp(-\lambda)$
$\mathrm{E}(X) = \lambda$	$\mathrm{var}(X) = \lambda$

Beta: $Be(\alpha, \beta)$	`_beta(arg, shape1 =` α`, shape2 =` β`)`
$\alpha, \beta > 0$	$0 < x < 1$
$\mathrm{p}(x) = \mathrm{const} \cdot x^{\alpha-1}(1-x)^{\beta-1}$	
$\mathrm{E}(X) = \frac{\alpha}{\alpha+\beta}$	$\mathrm{mod}(X) = \frac{\alpha-1}{\alpha+\beta-2}$ if $\alpha, \beta > 1$
For $\alpha = \beta = 1$ one obtains the uniform distribution on the interval (0, 1).	

Gamma: $Ga(\alpha, \beta)$	`_gamma(arg, shape =` α`, rate =` β`)`
$\alpha, \beta > 0$	$x > 0$
$\mathrm{p}(x) = \mathrm{const} \cdot x^{\alpha-1}\exp(-\beta x)$	
$\mathrm{E}(X) = \alpha/\beta$	$\mathrm{mod}(X) = (\alpha-1)/\beta$ if $\alpha > 1$

Normal: $N(\theta, \sigma^2)$	$\_\texttt{norm}(\texttt{arg}, \texttt{mu} = \theta, \texttt{sd} = \sigma)$
$\sigma^2 > 0$	
$\mathrm{p}(x) = \text{const} \cdot \frac{1}{\sigma} \exp\left(-\frac{1}{2}\frac{(x-\theta)^2}{\sigma^2}\right)$	$\mathrm{L}_x(\theta) \propto \exp\left(-\frac{1}{2}\frac{(x-\theta)^2}{\sigma^2}\right)$
$E(X) = \theta$	$\mathrm{var}(X) = \sigma^2$
$N(0,1)$ is called standard normal distribution.	

Log–normal: $LN(\theta, \sigma^2)$	$\_\texttt{lnorm}(\texttt{arg}, \texttt{meanlog} = \theta, \texttt{sdlog} = \sigma)$
$\sigma^2 > 0$	$x > 0$
$E(X) = \exp(\theta + \sigma^2/2)$	$\mathrm{mod}(X) = \exp(\theta - \sigma^2)$
$\mathrm{var}(X) = (\exp(\sigma^2) - 1) \cdot E(X)^2$	
If X is normal, i.e. $X \sim N(\theta, \sigma^2)$, then $\exp(X) \sim LN(\theta, \sigma^2)$.	

References

Altham PME (1969) Exact Bayesian analysis of a 2 × 2 contingency table and Fisher's "exact" significance test. J R Stat Soc B 31:261–269

Bayarri MJ, Berger JO (2004) The interplay of Bayesian and frequentist analysis. Stat Sci 19(1):58–80

Bayes T (1763) An essay towards solving a problem in the doctrine of chances. Philos Trans R Soc 53:370–418

Berger JO, Sellke T (1987) Testing a point null hypothesis: irreconcilability of P values and evidence (with discussion). J Am Stat Assoc 82:112–139

Bernardo JM, Smith AFM (1994) Bayesian theory. Wiley, Chichester

Besag J, York J, Mollié A (1991) Bayesian image restoration, with two applications in spatial statistics (with discussion). Ann Inst Stat Math 43:1–59

Boice JD, Monson RR (1977) Breast cancer in women after repeated fluroscopic examinations of the chest. J Natl Cancer Inst 59:823–832

Box GEP (1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). J R Stat Soc A 143:383–430

Breslow NE (1996) Statistics in epidemiology: the case-control study. In: Armitage P, David HA (eds) Advances in biometry. Wiley, New York, pp 287–318

Clayton D, Hills M (1993) Statistical models in epidemiology. Oxford University Press, Oxford

Clayton D, Kaldor J (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43:671–681

Cornfield J (1951) A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. J Natl Cancer Inst 11:1269–1275

Cornfield J (1966) Sequential trials, sequential analysis and the likelihood principle. Am Stat 20:18–23

Cornfield J (1976) Recent methodological contributions to clinical trials. Am J Epidemiol 104:408–421

Cox DR (2005) Principles of statistical inference. Cambridge University Press, Cambridge

Davison AC (2003) Statistical models. Cambridge University Press, Cambridge

Edwards W, Lindman H, Savage LJ (1963) Bayesian statistical inference in psychological research. Psychol Rev 70:193–242

Fisher RA (1934) Statistical methods for research workers, 3rd edn. Oliver and Boyd, Edinburgh

Goodman SN (1999a) Towards evidence-based medical statistics. 1: the P value fallacy. Ann Int Med 130:995–1004

Goodman SN (1999b) Towards evidence-based medical statistics. 2: the Bayes factor. Ann Int Med 130:1005–1013

Goodman SN (2005) Introduction to Bayesian methods I: measuring the strength of evidence. Clin Trials 2:282–290

Greenland S (2006) Bayesian perspectives for epidemiological research: I. foundations and basic models. Int J Epidemiol 35:765–775

Greenland S (2007) Bayesian perspectives for epidemiological research: I. regression analysis. Int J Epidemiol 36:195–202

Greenland S (2008) Introduction to Bayesian statistics. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 3rd edn. LWW, Philadelphia, pp 328–344

Greenland S, Rothman KJ (2008) Introduction to categorical statistics. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 3rd edn. LWW, Philadelphia, pp 238–257

Howard JV (1998) The 2 × 2 table: a discussion from the Bayesian viewpoint. Stat Sci 4:351–367

Liebermeister C (1877) Über Wahrscheinlichkeitsrechnung in Anwendung auf therapeutische Statistik. Sammlung klinischer Vorträge (Innere Medicin No. 31–64) 110:935–962

Mossman D, Berger JO (2001) Intervals for posttest probabilities: a comparison of 5 methods. Med Decis Making 21:498–507

Nurminen M, Mutanen P (1987) Exact Bayesian analysis of two proportions. Scand J Stat 14: 67–77

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) Uncertain judgements; eliciting experts' probabilities. Wiley, Chichester

Pawitan Y (2001) In all likelihood: statistical modelling and inference using likelihood. Oxford University Press, New York

Rothman KJ (2002) Epidemiology; an introduction. Oxford University Press, New York

Rue H, Held L (2005) Gaussian Markov random fields: theory and applications. Chapman and Hall, Boca Raton

Savitz DA, Wachtel H, Barnes FA, John EM, Tvrdik JG (1988) Case-control study of childhood cancer and exposure to 60-Hz magnetic fields. Am J Epidemiol 128:21–38

Sellke T, Bayarri MJ, Berger JO (2001) Calibration of p values for testing precise null hypotheses. Am Stat 55:62–71

Seneta E (1994) Carl Liebermeister's hypergeometric tails. Hist Math 21:453–462

Seneta E, Phipps MC (2001) On the comparison of two observed frequencies. Biom J 43:23–43

Spiegelhalter DJ, Abrams KR, Myles JP (2004) Bayesian approaches to clinical trials and health-care evaluation. Wiley, New York

Survival Analysis

32

Peter D. Sasieni and Adam R. Brentnall

Contents

P.D. Sasieni (✉) • A.R. Brentnall
Wolfson Institute of Preventive Medicine, Queen Mary University of London, London, UK

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_18,

32.1 Introduction

The term survival analysis originally referred to statistical study of the time to death of a group of individuals. From a mathematical perspective, it is irrelevant whether one is studying time until death or time to any other event, and so the term has come to be applied to methods for analyzing "time-to-event data." Although often not explicitly stated, we are always interested in the time between two events. For instance, one might be studying the age of death (the time from birth until death), survival of cancer patients (the time from diagnosis until death), or the incubation time of a virus (time from infection until the development of symptomatic disease). Survival analysis is more complicated than the analysis of other measurements because one often has only partial information regarding the survival time for some individuals. The most common form of partial information arises when a study is stopped before all participants have died. At that point, we might know that Mrs Patel survived for at least 3.7 years, but not know whether she will die a week later or 25 years later. The observation on Mrs Patel is said to be (right) censored at 3.7 years.

The goal of a survival analysis might be to describe the survival distribution for a group of individuals. One might wish to present the median age at onset of a particular disease and add that 90% of cases occur before a certain age. More often, epidemiologists are interested in factors that influence survival. In such instances, the aim of survival analysis is to estimate the effect of the factors on survival times.

Occasionally, survival times may vary little between individuals: the vast majority of humans is born at the end of 36 to 42 weeks of gestation. More frequently, survival times can vary hugely: the duration of a detectable viral infection could be anything from a few days to many years. In such circumstances, it is convenient to describe the rate at which the event (clearance of the virus) occurs. Using rates is particularly appealing when the event will never be observed in the majority of individuals being studied. It makes sense to talk about the average rate of breast cancer in a population even though the majority of the population will never get breast cancer. Studying disease rates and the factors that influence them is indeed central to much of epidemiology.

In this chapter, we consider three major objectives of survival analysis:

- The description of the survival experience of a group of individuals
- Comparison of the survival between two or more groups
- Regression analysis of variables that influence survival

We describe analyses appropriate for the most common study designs in epidemiology and briefly discuss methods that can be used in more complicated settings. The chapter is intended to give the reader a broad overview of the main techniques in survival analysis and to provide examples of the sorts of epidemiological problems that are amenable to survival analysis. There are now a number of texts on survival analysis written for a non-mathematical audience. Most are written either from the perspective of clinical trials or engineering. The book by Marubini and Valsecchi (1995) is written for a clinical audience and provides detailed worked examples. The book by Breslow and Day (1987) is written specifically for cancer epidemiologists but focuses on design and analysis of cohort studies rather than specifically on survival analysis. More recent books on survival analysis written primarily for medical statisticians are Hosmer and Lemeshow (1999) and Therneau and Grambsch (2000). The book by Selvin (2008) is specifically written for epidemiologists.

32.2 Data

The data from epidemiological survival analysis studies might be categorized into two types. The first are large data sets that can be analyzed at an aggregate level, even if the data are available at an individual level. Analyses of these data sets, perhaps obtained from a cancer registry, often focus on changes in rates of disease over time. The second sort of data are from smaller studies, and the analysis is carried out using individual-level data. Information about the characteristics of each person (e.g., his/her sex and blood pressure) is recorded with his/her survival data. Such data are commonly found in cohort studies. To provide a taste of the wide variety of survival data sets from epidemiological studies, we next present some examples. Their analysis will be described later on.

Example 32.1. Rates of cervical cancer
Data from different countries' cancer registries were extracted for the years 1982–1989 by Sasieni and Adams (1999). The aim was to compare the rate at which women in each country developed cervical cancer.

Example 32.2. Skin cancer in organ transplant recipients
Organ transplant recipients have an increased risk of skin cancer post-operation. Harwood et al. (2013) considered 931 patients in the UK who were transplanted with a kidney and were followed for more than 6 months. They analyzed the time to first and subsequent skin cancers by relating them to risk factors, in order to inform possible surveillance strategies.

Example 32.3. Cervical cancer in England and Wales since 1950
Cervical cancer mortality rates changed considerably during the second half of the twentieth century. It is well known that the rates vary with age, being essentially 0 until the age of 20 and then increasing slowly at first and rapidly in the 30s before reaching a plateau at about age 55. One explanation for the changing rates of cervical cancer over time is the cohort effect, whereby women's exposure to the human papilloma virus (HPV) during their teens and 20s largely determines their risk of cervical cancer throughout the rest of their lives. Thus, cervical cancer rates may be seen to vary as a function of the year of birth. Additionally changes in calendar time may affect the mortality rate from cervical cancer regardless of a woman's underlying risk of the disease. For example, improvements in treatment might reduce the mortality from cervical cancer at all ages. Sasieni and Adams (2000) used data from England and Wales to investigate age, birth cohort, and the secular trends that may be attributed to screening.

Example 32.4. Breast cancer screening
Lawrence et al. (2009) examined all breast cancers notified to the UK West Midlands Cancer Intelligence Unit and diagnosed in women aged between 50 and 74, from 1988 to 2004. The 26,766 cases were made up of 10,100 who were detected at screening and 15,862 with symptomatic diagnosis, of whom 6,009 women were diagnosed between screens (interval cancers) and 9,853 who had not attended screening. The endpoint used was death from breast cancer. The aim of the analysis was to compare the survival of screen-detected and symptomatic cases.

Example 32.5. Cohort of individuals exposed to medicinal arsenic
A cohort of 478 patients treated with Fowler's solution (potassium arsenite) between 1945 and 1965 were followed until the end of 1990 during which period 188 patients died (Cuzick et al. 1992). The endpoint of interest was death from cancer. Completeness of follow-up was achieved through flagging determined from the death certificates.

Example 32.6. Blood transfusions and AIDS
De Gruttola and Lagakos (1989) studied two groups of hemophiliacs who received blood transfusions after 1978 and before blood was screened for HIV infection. One group of 105 patients had a light treatment and received less of the blood factor; there were 157 who were heavily treated. The aim of the analysis was to compare the two groups' times of infection and symptoms.

Example 32.7. Assessment of survival following liver transplantation
Keiding et al. (1990) compared the survival experience of the first 38 primary biliary cirrhosis patients treated in the Nordic countries with liver transplantation with the directly standardized survival of 82 patients receiving transplants in England.

Example 32.8. Bone marrow transplant in childhood leukemia
Galimberti et al. (2002) considered the disease-free survival of 30 children with acute lymphoblastic leukemia (ALL) who were treated with allogenic bone marrow transplants while in first remission.

Example 32.9. Skin cancer
Leung and Elashoff (1996) considered data on $1,548$ patients with melanoma. The patients were followed periodically to detect changes in disease stage. The authors were primarily interested in the time from metastasis to death, but they also wanted to investigate whether this was dependent on the time from stage II disease to metastasis.

Example 32.10. Testing a treatment for primary biliary cirrhosis
Christensen et al. (1985) followed up 216 patients with primary biliary cirrhosis in a clinical trial, who were randomized to either azathioprine or a placebo. Twenty-five explanatory variables on the patient were also recorded. The primary endpoint of interest was death. Recruitment was over 6 years and follow-up a further 6 years. One hundred of the 216 patients died.

Example 32.11. Infant mortality
Moger and Aalen (2005) had survival data on $48,357$ infants from $24,077$ multiple births (e.g., twins, triplets) in Norway between 1967 to 2002 and examined all infants who survived the first week of life. There were 443 deaths, including 23 sibships with two deaths and one sibship with three deaths. The authors aimed to estimate the survival taking into account that the data might not be independent due to common genetic and environmental factors that affect siblings.

32.3 Concepts

This section covers some concepts that are helpful to analyze data such as described above, where the outcome measurement is the time to an event.

32.3.1 Walking Backwards into the Future

A key feature, perhaps the key feature of survival analysis, is that it relates to events that occur in time. In Western cultures, we tend to think in terms of looking forward into the future or backwards into the past. When thinking about survival analysis, it is useful to have the opposite picture. Imagine that you are walking backwards into the future so that the past is in front of you. You can look down to see things in the present and up to view things in the distant past. In order to look into the future, you would need to turn your head around, which is, alas, impossible. Rather the future is revealed as you walk backwards into it. Such a metaphor is important when deciding start and endpoints, which is considered further in Sect. 32.4.2.

32.3.2 Fundamental Quantities

In her first year, a gyneco-oncologist treated eight women with newly diagnosed cervical cancer. One died within 6 months, two more died over the next 3 years, and one died from a stroke 4 years later. Six years after she started, four of the patients were still alive. How should one summarize the survival times of the eight patients? Even with 800 patients and 20 years of follow-up, the problem is not trivial and there is no optimal solution, but some common quantities that help are presented next. Although some of this material may seem to be rather abstract and of little direct relevance to epidemiologists, it is necessary to have a rudimentary understanding of the different concepts in order to be able to critically assess studies analyzed using survival methods.

The random variable of interest in survival analysis is the time from an initiating event to a terminal event. There are several equivalent ways to describe the distribution of this random variable. Outside of survival analysis, one might define the distribution through its density or distribution function or choose to describe only certain features such as the mean and variance. With time-to-event data, it is more usual to define the distribution in terms of either the *survival function* $S(t)$ or the *hazard* or event-rate function $\lambda(t)$, where t is a chosen time after the initiating event. If time is discrete, so that for instance the event can only occur at $t = 1, 2, 3, \ldots$, then the hazard is the probability of an event at t, given survival to t. Most of this chapter is concerned with continuous time, where the survival is a real number greater than zero. For this setup, the hazard function is defined by considering the rate of an event in a short period of time after t (the probability of an event in a short period of time after t given survival to t divided by the length of the period), as the short period gets progressively smaller. In summary, if T is used to denote the continuous random survival time, then:

The Survival Function $S(t)$ is the probability that an event did not occur before t, that is, the event time T is greater or equal to t, and

The Hazard Function $\lambda(t)$ is the instantaneous rate of an event occurring at t, given that it has not yet happened by t.

For example, if interest is in time from contracting a disease to death, then given the time t since diagnosis, the survival function provides the probability of survival to date; the hazard function gives the rate of death at the current time t, given the individual is still alive. For this reason, the hazard function is sometimes called the force of mortality. In symbols

$$S(t) = \Pr(T \geq t) \quad \text{and} \quad \lambda(t) = \lim_{\Delta \to 0} \Pr(T < t + \Delta | T \geq t)/\Delta.$$

Thus, $\lambda(t)\Delta$ is the approximate probability of the event between t and $t + \Delta$; equivalently $1-\lambda(t)\Delta$ is the approximate probability of no event. One can therefore see a relationship between the hazard and survivor functions: $S(t) \approx S(t-\Delta)\{1-\lambda(t)\Delta\}$. That is, the survival function at t is approximately the survival function at t minus a small time Δ, multiplied by the approximate probability of no event in the short period.

Another fundamental quantity used in survival analysis is the cumulative hazard function $\Lambda(t)$, defined as the area under the $\lambda(t)$ curve, or integral of $\lambda(t)$, between 0 and t. By definition, it is also related to $\lambda(t)$ over a small time period Δ through $\lambda(t)\Delta \approx \Lambda(t+\Delta) - \Lambda(t)$. We can therefore rewrite the relationship between the hazard and survivor function as $S(t) \approx S(t-\Delta)[1-\{\Lambda(t+\Delta)-\Lambda(t)\}]$. Extending the argument using a series of short time intervals Δ between $0 = s_0 < s_1 < \ldots < s_n = t$, the survival function at t is given by

$$S(t) \approx \prod_{i=1}^{n} [1 - \{\Lambda(s_i) - \Lambda(s_{i-1})\}] \, .$$

This is useful for Kaplan–Meier estimation later on in Sect. 32.5.2 and is also one way that leads to a compact relationship between the hazard and survivor functions where

$$S(t) = \exp\{-\Lambda(t)\}.$$

32.3.3 Censoring

One of the advantages of working in terms of the hazard function is that it can easily be estimated even in the presence of *right-censored* data (provided the censoring is independent of the event of interest). Right-censoring is the term used to describe the situation in which some individuals are known to have survived for a particular length of time, but it is unknown when beyond that time they died (or will die). Formally, we can consider a death time and a censoring time for each individual. We observe only the smaller of the two times and the knowledge of which came first. In order to be able to interpret analyses of censored data, one usually needs to assume that the censoring time imparts no information regarding the timing of the event of interest other than the fact that the event of interest had not happened by the censoring time. Censoring individuals because they become too ill to attend follow-up clinic, for instance, would create problems because censored individuals will probably have greater mortality in the short-term than uncensored individuals.

An extreme example of censoring is given by the study of disease incidence from a disease registry. The annual hazard or incidence rate is estimated by the number of events during the year divided by the number of individuals (half way through the year) at risk of getting the disease. Ideally, the incidence rate should be calculated for a cohort identified at the beginning of the year, but this is rarely done by population registries. By using the number at risk, one is able to deal not only with individuals who die without getting the disease but also those healthy individuals who immigrate into the area covered by the disease registry. The immigrants are subject to delayed entry or *left-truncation*: we do not know about disease in immigrants before they arrive. Analysis is undertaken assuming that had they got the disease before immigrating they would never have been recorded in the registry.

32.3.4 Competing Risks

The last concept introduced is competing risks or event causes. In previous sections, we have assumed that there is only a single, possibly censored, event of interest. A common approach taken in survival analysis is to treat all censored times in the same way. In practice, there may be many different causes of censoring. These can lead to unrealistic independent censoring assumptions, and the competing risks might need to be considered explicitly in order to draw meaningful conclusions about the event of interest.

For example, when looking at cause-specific mortality, one needs to think carefully about the interaction between different causes of death. The interaction may be structural as in the case when one considers the effect of coding rules on the mortality rates from pneumonia (immediate cause) and cancer (underlying cause) or on cancer of the "uterus site not otherwise specified" and cancer of the "cervix uterus." In these examples, an increase in one may be linked to a decrease in the other. In other examples, certain causes of death will be related due to a common underlying risk factor. Smokers are at substantially increased risk of lung cancer, cardiovascular disease, and respiratory disease. Similarly, obesity can lead to death from a variety of causes. Despite these examples, many researchers treat different causes of death as if they are independent; they concentrate on a particular cause of death and treat all other deaths as independent censoring events. Such an approach is particularly tempting when one is interested in an event other than death: in estimating the age at natural menopause, one would probably treat death as an independent source of censoring. In general, however, the assumption of independence may not hold, and it is meaningless to ask what would happen in the absence of all other causes of death. The key is to concentrate on observable quantities and not to make unsubstantiated guesses at what might happen in the absence of a particular cause. Competing risk is discussed further in Sect. 32.8.2.

A popular approach in epidemiology to side-stepping a full-blown competing risks model is to consider the excess hazard or the relative survival function. Given a standard hazard λ_0 from a reference population, the excess hazard λ_e is the difference between the observed hazard λ and the standard hazard: $\lambda_e = \lambda - \lambda_0$. The relative survival function S_r that follows is $S_r = S/S_0$, where S_0 is the survival function from the reference population. It is useful to contrast the relative survival for patients with a particular type of cancer to the cause-specific survival. If the cause of death is well determined, one might expect the two quantities to be similar. The cause-specific survival (i.e., the survival treating death from other causes as independent censoring) might be bigger than the relative survival if patients who die *with* cancer are incorrectly recorded as having died *from* cancer. Differences will also exist if the population is heterogeneous and those who get (and survive) cancer are not representative of the whole population. For instance, breast cancer patients (particularly those that survive) tend to be of higher social class and therefore have lower rates of mortality from other causes; by contrast, lung cancer patients tend to be smokers and have higher rates of mortality from other causes. Methods for relative survival are considered further in Sect. 32.6.5.

Example 32.12. Mortality from other causes in cancer patients
Sasieni et al. (2002) considered the probability of not dying prematurely as a measure of the proportion cured of a particular cancer. They compared (1) the probability of not dying from the excess hazard before dying from background mortality, (2) the probability of not dying from the cause-specific hazard before dying from background mortality, and (3) the probability of not dying from the excess mortality of other causes before dying from background mortality. For breast cancer, these three probabilities (given as percentages) were 67%, 76%, and 98%, respectively. Thus women with breast cancer are more likely to die of something other than breast cancer than are women without breast cancer, but the excess mortality due to other causes is small. By contrast, the three percentages for women with lung cancer were 11%, 16%, and 60%. So, $100 - 16 = 84\%$ of women with lung cancer die of it, and $100 - 60 = 40\%$ of those who don't still die prematurely.

32.4 Study Design

In this section, we consider some important ingredients for planning or critically assessing a study. The data from pooled, cohort, and period approaches are described, followed by the importance of choosing appropriate start and endpoints and the baseline group to make comparisons. The section ends with a brief overview of some designs covered elsewhere in this handbook that might be useful for survival analysis studies.

32.4.1 Survival Experience

Suppose that the aim of a study is to examine survival over a 20-year period. Then, three fundamental approaches to gathering data for survival analysis are:

1. *Pooled*. Use all patients diagnosed over a 20-year period.
2. *Cohort*. Use a cohort of patients, and follow them up for 20 years.
3. *Period*. Use patients who died in a particular (recent) period and who were diagnosed up to 20 years previously.

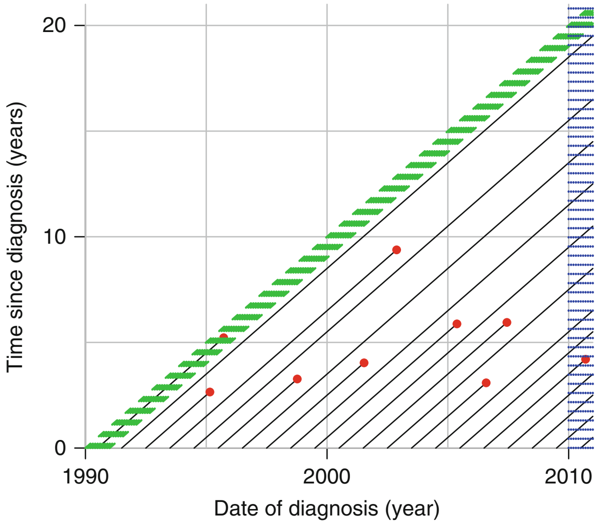


Fig. 32.1 A Lexis diagram to demonstrate some features of epidemiological survival data sets. On the x-axis is the calendar time that an individual had his/her initiating event. The y-axis shows his/her survival time: a line grows at a 45° angle until the terminating event shown by a *dot*. To explore survival, one might (i) pool all events, (ii) follow a cohort, or (iii) use the deaths from a recent period. To demonstrate, the first examines all the individual *lines* plotted, the second follows people from 1990 to 2010 (the *diagonal dashed bar*), and the third examines all the events in 2010 (*dashed vertical bar*)

A pictorial representation of these design options is in Fig. 32.1. The pooled method uses all available data, but if survival has improved over calendar time, the survival function will have been estimated from a mixture of patients given old and new treatments. The cohort approach has the clearest interpretation since it describes the survival experience of a particular cohort. However, using it to estimate 5-year survival will yield outdated estimates if survival has changed over the 20 years since members of the cohort were diagnosed. The period method will in many circumstances provide the best prediction of the current survival experience without explicitly modeling changes in survival over time (year of diagnosis). It is possible that the estimated survival might be inappropriate, if for instance a new treatment increases the hazard of dying in the first year but increases the proportion cured of the disease. More often the use of the most recent data to estimate the relevant hazard will mean that the period method will be favored. It has always been favored by demographers. It was explicitly introduced into epidemiology by Brenner and Gefeller (1996) and has been used by Brenner (2002) and by Sasieni et al. (2002) to study the long-term survival of cancer patients. Applying the period approach to data from the United States of America, Brenner (2002) estimated the 20-year relative survival for all types of cancer to be 51% compared to 40% using the cohort-based survival. The difference reflects the substantial improvement in survival over the past two decades.

32.4.2 Start and Endpoints

Survival analysis is the study of the time from an initiating event to a terminating event. Two aspects of event definitions are considered: (i) they cannot be defined using future data, and (ii) the study objective.

Event Definition The idea of walking backwards into the future is useful to decide what is an acceptable definition of an event. For instance, clinicians are fond of defining a patient as being in remission if he/she has been free of symptoms for a certain number of months (3 say). This is fine. What is not permissible from a probabilistic perspective is to then claim that remission started at the beginning of the 3-month period. As we walk backwards through time, we need to be able to see whether or not a patient is in remission without turning around to look into the future. The difficulty is that if someone has been symptom free for 2 months, we don't know whether or not he/she is in remission – that depends on what happens in the future. Practically, this causes problems in a survival analysis when there are patients in limbo: free from disease, but not for 3 months.

A light-hearted example of a poorly defined initiating event is given by the duration of pregnancy. It is impossible to be just 1 week pregnant! Since the duration of pregnancy is measured from a woman's last period but conception cannot occur until after ovulation (which is generally about 14 days after the previous period), 1 day after conception a woman is said to be 15 days pregnant.

Study Objective The study aim plays an important role in the choice of start and endpoints. When disease is not common in the target group, then death might not be a feasible endpoint: the size of sample required may be prohibitive, and so an earlier stage in the disease might be more appropriate. However, when disease is common among the target group, then death from any cause, sometimes called overall survival or all cause mortality, might be feasible.

There are benefits and drawbacks from choosing overall survival versus a cause-specific endpoint. Consider a study to examine survival in patients with a particular form of cancer (e.g., breast cancer). A drawback of overall survival is that it would dilute treatment effects by including death from causes that are not related to the cancer of interest, and so more data would be required to detect a difference than would be required using an endpoint of death from breast cancer. Indeed such primary endpoints are sometimes preferred (Cuzick 2008). An advantage of overall survival is that it avoids issues associated with cause-specific endpoints that arise from competing risks. For example, a meta-analysis of the early trials of radiotherapy and breast cancer showed that women treated with radiotherapy had lower rates of death from breast cancer but higher rates of death from cardiovascular disease (Cuzick et al. 1994). Later analyses showed that radiotherapy (particularly of the left breast) caused damage to the major blood vessels surrounding the heart. In this case, an analysis that focused solely on death from breast cancer and treated other events as censoring could be disastrously misleading. Similarly, an analysis that focused only on all-cause mortality would have missed the benefits and harms of radiotherapy. Understanding both has led to refinements in radiotherapy that maintained its therapeutic impact but minimized its adverse effects on the heart and major blood vessels.

32.4.3 Baseline Groups

Baseline groups are chosen for comparisons or as a way to standardize rates. The choice of baseline should depend on the study objectives and the nature of the cohort of interest. When studying mortality in a healthy cohort, it is usual to compare their survival experience to that of the general population, but when studying a cohort of very sick individuals or healthy individuals for an event that is not routinely registered, it may be desirable to compare their survival with a different control group. For instance, to assess whether a special residential environment helps to reduce suicide in schizophrenic patients, it would be better to compare them against patients cared for in the community than the general population. Similarly, when comparing relative survival in cancer patients by deprivation, one should preferably use deprivation-specific data on all-cause mortality in those without cancer. Some methods to compare survival of a cohort with a chosen baseline group are described in Sect. 32.6.2.

32.4.4 Choice of Design

Putting together chosen survival experience data, primary endpoints, and baseline groups into a suitable sampling schemes requires care. We review some common examples next that are also covered in more detail elsewhere in this handbook.

Matching A technique used in cohort studies to obtain a comparable survival curve is *matching* to identify patients from the baseline cohort, such as is done in case-control studies. Matching is usually undertaken once an event, such as the diagnosis of disease, has happened. If this event is the endpoint, such as in the time from birth to diagnosis, then the study has a period form. If the event is the start of the period of follow-up, such as in the time from diagnosis to death, then the study will have a cohort or pooled form. The basic idea is to match the cases to controls by using factors that are related to the hazard function. It is common to match on age and sex because the hazard function often increases with age, and men might have a different risk than women. However, the choice of variables used should depend on the study, as the following example demonstrates.

Example 32.8. Bone marrow transplant in childhood leukemia (cont.)
Matching was carried out at diagnosis of leukemia. Controls were selected from 397 ALL patients treated with chemotherapy. Matching was done on white blood cell count at diagnosis (0–10,000; 10,000–50,000; 50,000–100,000; >100,000/*cubic millimeter*), age at diagnosis (<1; 1–10; >10 years), immunophenotype (T-lineage; B-lineage), as well as clinical center and front-line chemotherapy protocol. Additionally, controls had to survive

in remission at least as long as the time from remission to transplant in the patient that they matched. Each of the 30 transplant patients was matched to between 1 and 7 controls. In all, there were 130 controls and no control was matched to more than 1 transplanted patient.

Sampling from the Risk Set Large cohort studies with long-term follow-up are extremely expensive. Studies requiring detailed analysis of food diaries or complicated analysis of blood, urine, or tissue can be prohibitively expensive. For this reason, many epidemiologists have set up cohort studies collecting questionnaires, sera, or urine and storing them for later use. The idea is to nest a case-control study within the cohort study (Langholz and Goldstein 1996; cf. chapter ▶Modern Epidemiological Study Designs of this handbook).

As is well known, in a matched case-control study, it is rarely efficient to collect more than about four controls per case (Breslow and Day 1980). So in a cohort study, one might wish to select a few "controls" from the cohort each time an event (case) occurs. Another approach is to select a small sub-cohort at the beginning and to supplement it with all the cases that develop during follow-up of the main cohort. There are a number of complications related to whether one can use controls selected for one case as controls for later cases (assuming that they are still at risk) and what happens if a control later becomes a case, but such nested case-control and case-cohort designs are extremely important in very large cohort studies looking at diet, genetics, or molecular measures of exposure.

The idea of the nested case-control design was first put forward by Thomas (1977) and was considered at length by Prentice (1986a) and Borgan and Langholz (1993). The case-cohort design was proposed by Prentice (1986b) and has been studied by Self and Prentice (1988) and by Chen and Lo (1999). More recent research has focused on how best to select controls (Langholz and Borgan 1995; Borgan and Olsen 1999). Chen (2002) discussed how to fit a Cox model to a study in which crude covariate information is available on the entire cohort and complete covariate information is available on a sample, but he assumed that the validation sample is chosen at random and it is not applicable to nested case-control designs.

Sampling Bias Finally, it is important to acknowledge sampling biases in the design before analyzing the data.

Example 32.4. Breast cancer screening (cont.)
Analysis of the difference in survival between screen-detected and symptomatic diagnosis is subject to two main sources of bias. Firstly, the initiating event is different (lead time bias). Secondly, slow growing tumors are more likely to be detected by screening (length bias). Both issues are discussed further in this handbook (chapter ▶Screening of this handbook).

Example 32.13. Immortal time bias
Suissa (2008) described observational studies from databases of general practitioner's prescription records that have shown significant beneficial effects of certain prescriptions. These studies ignored the issue that individuals who are exposed to the treatment necessarily survive until it is prescribed: before prescription, they are effectively immortal. Unless immortal time is acknowledged in the analysis, the treatment effect will be biased in favor of the prescription. For a similar reason, on average parents live longer than their children,

but this does not mean that you will live longer if you have children as an adult. It is because anyone who is a parent could not have died as an infant.

32.5 Descriptive Measures

Armed with data from a suitably designed study, the next stage is to summarize it. Some common approaches in survival analysis are first motivated by comparing them to common approaches for data without censoring or the usual univariate summaries. The survival analysis summaries are then discussed in detail.

32.5.1 Univariate Survival Analysis

When the data are not censored, certain univariate statistics and plots are often applied to summarize them. Possible replacements are presented in Table 32.1 and Fig. 32.2, and a more detailed description follows.

Table 32.1 Summary statistics for univariate analysis without censoring and possible replacements for survival data with censoring. The replacements marked * are obtained from Kaplan–Meier or life-table estimates of the survivor function. They are also shown in Fig. 32.2

Complete data	Survival data
Mean	Average rate
Histogram	Kaplan–Meier (KM) plot
Median	Median*
Interquartile range	Interquartile range*
% Improved (binary data)	%5-year survival*

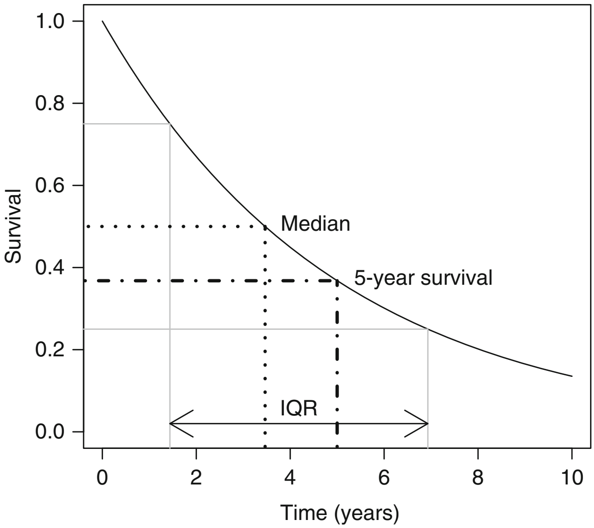


Fig. 32.2 A survival curve with the median, interquartile range (*IQR*), and 5-year survival statistics in Table 32.1 highlighted. Thus about half of the individuals survive more than 3.5 years (median), and a little over 35% survive at least 5 years (5-year survival). The survival for the middle 50% of individuals is between around 1.5–7.0 years (IQR)

In the presence of censoring, or when not everyone will experience the terminating event, the mean survival time may be impossible to estimate. However, a possible replacement is suggested by the fact that when there is no censoring, then the sample mean survival time is the inverse sample rate. This can be obtained when there is censoring.

Histograms are often used to show the spread of the data, but these do not transfer easily to survival data because of censoring. An equivalent way to show the data from a histogram is through a cumulative distribution function (*CDF*). This plot tends to be underused but can be more useful than a histogram, especially when different groups may be compared within the same plot. The survivor function is 1 minus the distribution function, and since it may be estimated (Kaplan–Meier), it can be used as replacement for the histogram.

If there are sufficient uncensored data, survival estimates provide a route to another summary of the location of a distribution, the median. They may also be used to gauge the spread of data through estimates of the interquartile range. Survival estimates are often obtained via Kaplan–Meier estimation, or life tables and these are covered in more detail in the following sections.

In some applications, the outcome is a binary indicator, perhaps denoting success or failure. A possible replacement for the success proportion with survival data is to fix the survival time to 5 years, say, and give an estimate of the proportion of cases that survived: the 5-year survival.

In the rest of this section, we consider how to estimate the survivor function to obtain some of the quantities and then discuss sensible ways to present the average rate.

32.5.2 Kaplan–Meier Survival Estimate

The product definition of the survival function in Sect. 32.3.2 leads naturally to the definition of the product-limit or Kaplan–Meier (1958) estimator of the survival function (see also Example 32.14).

Let $t_1 < t_2 < \ldots < t_n$ be distinct ordered event times observed in a cohort. Let d_j be the number of events at t_j and n_j the number in the cohort "at risk" of having an event observed at t_j. Then d_j/n_j will be a rough estimate of the probability of an event between t_{j-1} and t_j among those at risk at t_{j-1}. When the event of interest is death, once an individual has had an event, he/she is no longer at risk for further events. Similarly if an individual is only followed for 2 years, he/she is not at risk for an *observed* event beyond 2 years. More formally, the cumulative hazard between t_{j-1} and t_j is estimated by d_j/n_j, and $(n_j - d_j)/n_j$ estimates the probability of an individual at risk at t_{j-1} not dying by t_j. The Kaplan–Meier estimate of $S(t)$ is then the product over all t_j less than or equal to t of $(n_j - d_j)/n_j$. That is,

$$\hat{S}(t) = \prod_{t_j \le t} (n_j - d_j)/n_j = \prod_{t_j \le t} (1 - \hat{\lambda}_j).$$

So, in particular, the estimate at t_j is given by $(n_j - d_j)/n_j$ times the estimate at t_{j-1}.

Example 32.14. Cancer data from a clinical trial
To demonstrate the measures to summarize survival, we use the "cancer" data set provided by the statistical software Stata: There are 48 patients who received one of three different drugs, and the endpoint is death. Thirty-one patients died during follow-up.

A summary table of the data is shown in Table 32.2. Figure 32.3 presents a plot of the Kaplan–Meier survival estimates in each of the three arms.

Nelson–Aalen Cumulative Hazard Estimate Some authors prefer to present data in terms of the cumulative hazard function $\Lambda(t)$. The Nelson–Aalen estimator is given by

$$\hat{\Lambda}(t) = \sum_{t_j \le t} d_j/n_j = \sum_{t_j \le t} \hat{\lambda}_j .$$

In words, it is the sum of the estimated hazards at each event time. It is related to the Kaplan–Meier estimate because the Kaplan–Meier estimate uses the same hazard estimate $\hat{\lambda}_j$. The cumulative hazard can sometimes be a more useful way to assess

Table 32.2 Descriptive survival in the three treatment arms

	Survival at 1 year	Survival at 2 years	Median survival (years)	Mean annual mortality rate
Placebo ($n = 20$)	0.23	0	0.67	1.27
Drug 2 ($n = 14$)	0.85	0.21	1.83	0.34
Drug 3 ($n = 14$)	0.86	0.77	2.75	0.21

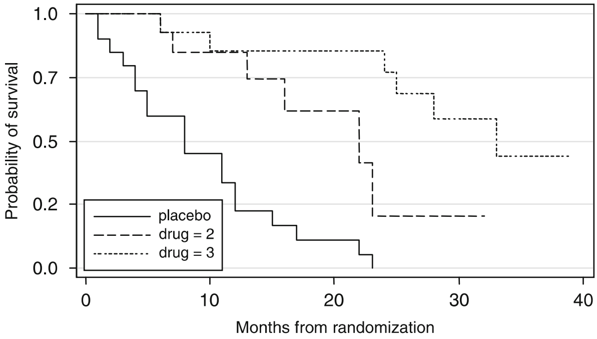


Fig. 32.3 Kaplan–Meier estimates of the survival of cancer patients over 3 years. The three estimates correspond to patients on different drugs. The chart shows an estimate of the probability of survival to each point on the x-axis. It starts at 1.0 on the y-axis because everyone is alive when they entered the trial. Given enough follow-up time, all the survival curves would end at zero; here they are plotted until the longest observation in each arm

how the hazard changes than the survival function. It is particularly useful when an individual may have more than one event.

Example 32.2. Skin cancer in organ transplant recipients (cont.)
The Nelson–Aalen cumulative hazard estimate was useful for this analysis because it was very common to get a second cancer after the first one. Around one in three developed at least one cancer per year following a first cancer, and almost one in six got more than one per year. This showed that regular surveillance was needed once a first skin cancer had been diagnosed.

32.5.3 Life Tables

We have seen that a survival function can be estimated by the product-limit or Kaplan–Meier estimate. Such an approach is standard for small and moderate cohorts. For larger cohorts and for aggregate data, it is more usual to estimate the hazard for a year at a time and to use the hazard function to estimate survival. Sometimes people calculate the person-years at risk in a given time period exactly and then use that as the denominator for calculating the rate. Often, however, the person-years at risk is approximated by multiplying the length of the interval by the mean of the number of people at risk at the beginning of the interval and the number at risk at the end of the interval. Formally one is assuming that on average, anyone who "drops out" or "enters" during the interval will have been at risk for half of the interval. Dividing the number of events in an interval by this estimate of the person-years leads to the life-table estimate of survival. The probability of surviving a long period is computed by multiplying the conditional probabilities of surviving each of the intervals constituting it. Suppose the conditional probabilities have been estimated in intervals of width one and that we wish to estimate the survival at time t such that j is the largest interval that is less than or equal to t. Then the $S(t)$ is estimated by the product of $(1 - p_i)$ for $i < j$ multiplied by $(1 - p_j)^{t-j}$. This method has an extremely long history being first devised in the seventeenth century by Edmund Halley to describe the mortality of the people of Breslau (Halley 1693). The probabilities p_i are calculated by the formula $d_i/(n_i - c_i/2)$ where n_i is the number at risk at the start of the ith interval, d_i is the number dying, and c_i is the number censoring during ith interval.

Whereas in clinical studies, it would be standard to base life tables on a group of patients followed from diagnosis to death; actuarial life tables are not constructed by observing a cohort of newborns until the last survivor dies. Rather they are based on estimates of probabilities of death, given survival to various ages, derived from the mortality experienced by the entire population over a few consecutive years. In that case, there is little point in noting the numbers censored in each interval; rather, one needs to keep track of the size of the risk set. The following fictitious example is provided to show the calculations involved (Table 32.3). Initially 2,456 patients are followed. In the first year, 543 (22%) die leaving 1,913 patients at the beginning of the second year. In the second year, 265 patients die and 22 are censored. In the third year, 934 new patients enter the study. (This could be due to transfer from

Table 32.3 Fictitious life table

Years from diagnosis	No. at start of interval	No. censored	No. of new patients in interval	Deaths	Conditional probability of death	Cumulative probability of survival
j	n_j	c_j		d_j	p_j	S_j
0	2,456	0	0	543	0.2211	1.0000
1	1,913	22	0	265	0.1393	0.7789
2	1,626	130	934	302	0.1489	0.6704
3	2,128	219	0	321	0.1590	0.5706
4	1,588	336	971	317	0.1664	0.4798
5	1,906	447	0	278	0.1652	0.4000

another hospital, and we assume that there is no information on the numbers treated from diagnosis in that hospital but that 934 who were diagnosed between 2 and 3 years earlier were transferred.) The conditional probability of dying in the third year is calculated as $302/(1{,}626 - 130/2 + 934/2) = 0.1489$. Thus the estimated probability of surviving until the beginning of year 4 is $(1-0.2211) * (1-0.1393) * (1 - 0.1489) = 0.5706$.

32.5.4 Standardizing Rates for Age

Since the rates of many diseases are highly age dependent, it may be misleading to present the crude disease rates. We consider three approaches to standardize the rates for age:

Directly Standardized Rate. The traditional approach (cf. chapter ▸Descriptive Studies of this handbook) takes a weighted average of the age-specific rates λ_i, with weights chosen to be proportional to the numbers p_{0i} in each age group in a standard population. This estimate ($\sum \lambda_i p_{0i} / \sum p_{0i}$) is known as the *directly standardized* rate. An alternative approach, indirect standardization is introduced in Sect. 32.6.2.

Cumulative Rate. Another approach is to use the cumulative rate up to some age such as 74 complete years (Day 1976). This has the advantage of dispensing with the selection of a standard population, and it is straightforward to convert from the cumulative rate to the cumulative risk. If the cumulative rate is 1 in x, then, to a very close approximation, the cumulative risk is 1 in $x + 0.5$ (because the cumulative risk, or survival function $\exp(-1/x)$, is mathematically approximately the same as $1/(x + 0.5)$).

Lifetime Risk. A third, less used, approach is to take a weighted sum of the age-specific rates using standard weights (corresponding to the probability of living to that age) so that the cumulative rate has the interpretation of the lifetime risk in a population with standard mortality rates (Sasieni and Adams 1999). An issue

Table 32.4 Cervical cancer rates in various populations around the world

	Age-standardized rate per 1,000 (world standard)	Cumulative rate (from birth) to age 74 per 1,000	Lifetime risk per 1,000
Cali, Colombia	49	50	71
Trujillo, Peru	55	58	77
USA (SEER)			
White	7	7	8
Black	12	12	15
Israel			
Jews	4	4	5
Non-Jews	3	3	3
Denmark	16	16	17
Finland	4	5	6
England and Wales	12	12	13
New Zealand			
Maori	30	31	34
Non-Maori	12	12	13

that sometimes arises is that a single person can have multiple registrations, for example, one for breast cancer and another for colorectal cancer. When data on first registrations alone are not available, Sasieni et al. (2011) developed an adjustment.

Example 32.1. Rates of cervical cancer (cont.)
Table 32.4 extracted from Sasieni and Adams (1999) compares three measures of cervical cancer rates from various cancer registries for the years 1982–1989. The lifetime risk is calculated using all-cause mortality rates from England and Wales in 1992 as the standard. It is seen that the numerical difference between the various measures is not great, but the lifetime risk has perhaps the most natural interpretation.

The age-standardized rate is an average rate for the population (including those for whom the rate is essentially zero, e.g., children). It will not correspond to the crude rate unless the age distribution of the population is identical to that of the standard population. The cumulative rate to age 74 is more easily interpreted, but most people will think of it as a probability, and this is not quite correct – the difference will be particularly great when the cumulative rate is large (or in the presence of competing risks – see Sect. 32.8.2). The lifetime risk can be directly interpreted as the probability of being diagnosed with cervix cancer. It uses the hazard of all-cause mortality from a standard population and assumes that cervical cancer incidence rates are independent of mortality rates from other causes (see competing risks).

32.6 Comparison Between Groups

This section shows some methods to compare the survival experience of two or more groups, where the aim is to quantify the difference. The exponential distribution is first introduced as a way to make inferential statements. Subsequent sections comment on additional ways to compare the groups depending on the study's baseline comparison group.

32.6.1 Exponential Distribution

Model In many applications in epidemiology, it is reasonable to assume that the hazard function is constant over short intervals. The family of distributions with a constant hazard rate is known as the *exponential distribution*, and a distribution with a hazard function that is constant on intervals is known as a piecewise exponential distribution. The likelihood for such piecewise exponential models can be written quite simply and should be the basis for statistical analysis. Confidence intervals and inferential statements follow in the usual way.

Use with Aggregate Data The exponential distribution is often used to test for differences in aggregate data and in life tables. The model parameter is obtained by dividing the number of events in an interval by the number of person-years at risk, which is calculated by adding up the length of time that each individual was at risk within the interval. For instance, inference might be required for cancer incidence data given in 5-year age bands or for survival data given in terms of the number of whole years from diagnosis.

Checking the Assumption A simple way to check whether the constant hazard is reasonable is to use a Nelson–Aalen plot (Sect. 32.5.2) over the period thought to have a constant hazard. Since the model specifies that $\lambda(t) = \lambda$, we know that $\Lambda(t) = \lambda t$. If the exponential distribution is a reasonable approximation, then the Nelson–Aalen estimate of $\Lambda(t)$ should look like a straight line, with slope λ.

Poisson Approximation Although inference using the exponential distribution is relatively straightforward, a Poisson distribution is sometimes applied instead. This is a reasonable approach when one has a large number of individuals all of whom are at risk in an interval with a constant hazard λ per year, then the likelihood looks like a Poisson likelihood: the number of events observed will (to a close approximation) follow a Poisson distribution with mean λn, where n is the number of person-years at risk in the interval. The approximation will be good provided the number of individuals at risk in the interval is at least 10 times λn.

32.6.2 Standardized Mortality Ratios

The survival experience of a group of individuals relative to another group has been described using standardized mortality ratios for over 200 years (Keiding 1987), having first been described by English actuaries in the 1780s. The idea is to compare the expected number of deaths in two groups. We next describe standardized mortality ratios when a cohort is compared with an external baseline group such as the general population, then consider comparisons within a cohort, and provide a way to use the measure when the cohort are all healthy at the start of a study.

Standardization with External Baseline Groups are defined by age and sex and possibly the year of risk. For each group i, let λ_{0i} and n_{0i} denote the mortality rate and the number of person-years at risk in the reference population and λ_i and n_i be the corresponding quantities in the study cohort. The standardized mortality ratio (*SMR*) is defined as SMR $= \sum \lambda_i n_i / \sum \lambda_{0i} n_i$ and is equal to observed number of deaths in the study cohort divided by the expected number under the assumption that the reference population rates apply (cf. chapter ▸Descriptive Studies of this handbook).

Corresponding to the SMR, the indirectly standardized rate is the SMR times the standardized rate in the reference population: SMR $* \sum \lambda_{0i} n_{0i} / \sum n_{0i}$. Please note that the reference population does not serve as a "standard" population for the calculation of an SMR, that is, for the indirect standardization. Rather, the mortality rate of the reference population is standardized by the age distribution of the study cohort. This should be contrasted with the directly standardized rate: $\sum \lambda_i n_{0i} / \sum n_{0i}$. The latter is more widely used when the study cohort is very large but requires reasonable estimates of all the individual λ_i and should be avoided unless all the n_i are reasonably large.

Conditional Inference within a Cohort In cohort studies, one often wants to compare the number of events in two groups within a cohort. This is done by considering the observed number of events and the "expected" number of events in each group. The expected number is calculated as in the SMR by $\sum \lambda_{0i} n_i$. Statistical inference is usually based on the assumption of proportional hazards. That is, it is assumed that in each group, $\lambda_i = k\lambda_{0i}$ for all i. Suppose the observed and expected number of events are O_1 and E_1 in group 1 and O_2 and E_2 in group 2, respectively. Then, under the null hypothesis of equal hazards in the two groups ($\lambda = 1$), O_1 is distributed as a binomial sample from a population of size $O_1 + O_2$ with probability of $E_1/(E_1 + E_2)$. Hence the null hypothesis of no difference between the groups can be tested using an exact binomial test.

Example 32.5. Cohort of individual exposed to medicinal arsenic (cont.)
A comparison was made between the observed number of deaths from various causes (before the age of 85) and the expected number using age-, sex-, and calendar year-adjusted rates from England and Wales. Treating the cohort as a whole, SMRs were calculated for various causes of death (Table 32.5). There was a slight (and non-significant) deficit of death overall (suggesting that the cohort is slightly healthier than the population as a whole). The observed number of five deaths from bladder cancer was three times greater than expected and represented a significant increase compared to the general population.

Of the 5 bladder cancer deaths, 4 were in individuals with a cumulative dose of over 500 milligram (mg). The expected numbers were 0.83 for those exposed to less than 500 mg and 0.80 for those exposed to over 500 mg. Thus to test whether the risk was greater at the higher dose, one calculates the binomial probability of four or more "successes" out of five with a probability of 0.49 ($= 0.80/(0.80 + 0.83)$) each. The binomial probability is 0.18. Hence although the tendency was for the individuals who died from bladder cancer to have been exposed to a greater dose of arsenic, the association was not statistically significant.

Deaths in an Initially Disease-Free Cohort In most applications of the person-years method, it is simply a case of multiplying the number of years at risk by

Table 32.5 Mortality in a cohort of 478 individuals exposed to medicinal arsenic

	Observed number	Expected number	*SMR*	95% CI
All causes	188	209.1	0.90	0.77–1.03
All cancers	47	49.3	0.95	0.70–1.3
All circulatory diseases	97	106.5	0.91	0.74–1.1
Bladder cancer	5	1.6	3.07	1.01–7.3

SMR standardized mortality ratio; *CI* confidence interval

the population risk (mortality rate of a particular disease) and adding these to get an expected number of events (disease-specific deaths). Such estimates often overestimate the true mortality in a given study cohort because of what is known as "the healthy worker effect." Cohorts of individuals working in a particular occupation will often have low mortality rates for the first few years of follow-up simply because if they are working, then they are probably healthy. A particular problem arises in cancer screening studies in which one excludes anyone who already has cancer from the study. For such a cohort, it is not appropriate to simply apply the cancer-specific mortality rates because individuals are all cancer-free at entry. Instead one should explicitly model the probability of first getting cancer (using population incidence rates) and then dying from cancer (using cancer survival rates). Explicit formulae are given in Sasieni (2003), and the method is applied to evaluating colorectal cancer mortality in a cohort having colonoscopy by Dove-Edwin et al. (2005).

32.6.3 Comparisons That Use Matching

When the sample has been obtained by matching, such as in case-control studies, the Kaplan–Meier survival estimate is a useful way to compare the groups either directly or via a summary statistics from Table 32.1. With one-to-one matching, one could simply use the two Kaplan–Meier curves to obtain the statistics or plot the curves directly with standard error bars. If a variable number of "controls" are selected for each patient in the "special cohort," one needs to produce a weighted Kaplan–Meier curve. Here one simply weights each control by the inverse of the number of controls in the matched set so that each matched set of controls has weight 1. Estimation of the survival curve is straightforward, but estimation of its variance is more complicated (Winnett and Sasieni 2002).

Example 32.8. Bone marrow transplant in childhood leukemia (cont.)
Potential controls not matching to any transplanted patients tended to have better survival, although they also included patients who died before there was time for a transplant. Similarly, when there were multiple matching controls, their survival tended to be better. Thus, compared to all 397 controls, the matched sample of 130 controls had excellent survival for the first 6 months, but the survival curves crossed within 12 months of remission, and thereafter the smaller group had worse survival. After taking account of the variable number of controls, the weighted Nelson–Aalen estimate showed that the survival in transplant patients compared with that in those treated was even better (Fig. 32.4).

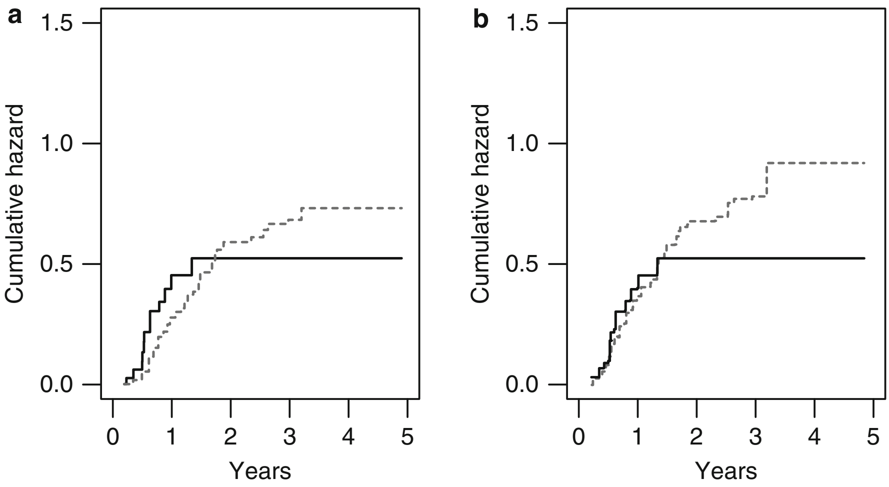


Fig. 32.4 Estimates of the cumulative hazard for the bone marrow transplant (BMT) patients (*solid lines*) and the conventionally treated (CT) patients (*dashed lines*). (**a**) Standard estimates based on 30 BMT and a matched group of 130 CT patients. (**b**) Standard estimate for the 30 BMT patients and weighted estimate for the 130 matched CT patients. In both cases, the CT patients have higher cumulative hazard beyond 2 years

32.6.4 Inference Without a Reference Population

The problem of comparing survival without the use of a reference population will be most relevant when the mortality of all groups in the study is substantially greater than that of the general population so that one can almost ignore the "background mortality." That is, one just compares the groups themselves.

A graphical solution is achieved by plotting the estimated survival function (using the Kaplan–Meier or the life-table approach) on a common set of axes. Comparison of survival at a fixed time (e.g., 3 years) can be made using Greenwood's formula for the variance of the survival estimate (Greenwood 1926). Greenwood's estimate is widely available in software packages, and a formula for calculating the estimated variance can be found in most texts on survival analysis. More sophisticated confidence bands, for the whole survival function, have been proposed by Hall and Wellner (1980).

Another useful summary of the survival in each group is the average event rate, which is estimated by dividing the number of events by the total person-years at risk. For instance, 12 deaths in a group of 100 patients who were followed for a total of 482 person-years would yield a rate of 12/482 or 2.5% per year. This simple approach is also amenable to hypothesis testing. Formally one is assuming a constant rate or exponential model, and the problem of testing between two groups is equivalent to testing for the equality of rate parameters in an exponential regression model.

Table 32.6 2×2 table formed at time T_j

	Die at T_j	Survive T_j	At risk at T_j
Group 1	d_{1j}	$n_{1j}–d_{1j}$	n_{1j}
Group 2	d_{2j}	$n_{2j}–d_{2j}$	n_{2j}

The test statistic most often used in medical statistics is a variant of the log-rank test first proposed by Mantel (1966). This is a test that is intended for use with censored survival data (including right-censoring and left-truncation) and which is most powerful when the proportional hazards model holds. The proportional hazards model is that the hazard in each group g is given by $\lambda_g(t) = \theta_g \lambda_0(t)$. In other words, the ratio of the hazard functions is constant over time. Readers, who are familiar with the Mantel–Haenszel test for a series of 2×2 tables, might like to know that the log-rank test for two groups is equivalent to the Mantel–Haenszel test applied to the set of 2×2 tables, one for each event time. Consider the situation at each event time T_j. Let d_{ij} be the number of events in group i at time T_j and n_{ij} the number at risk as illustrated in Table 32.6.

Although these 2×2 tables are not independent, they are conditionally independent (given the marginals), and the usual formula for the variance of the Mantel–Haenszel test statistic holds.

In epidemiology, it is often useful to be able to stratify the log-rank test so that a valid test can be obtained after adjusting for some other factor. Formally, the test assumes that within each stratum, proportional hazards hold with the same constants of proportionality: $\lambda_{gs}(t) = \theta_g \lambda_{0s}(t)$ for each stratum s and each group g.

Example 32.14. Cancer data from a clinical trial (cont.)
The log-rank statistic (which, under the null hypothesis of no differences in effect on survival between the three drugs, is asymptotically distributed as chi-squared with 2 degrees of freedom (df)) was 30.2 ($p < 0.0001$). Just comparing drugs 2 and 3, gave a log-rank of 3.4 (1 df, $p = 0.065$). However, it was noted that older patients fared worse – the log-rank test for age under 55 versus age over 55, stratified for treatment, had $p = 0.007$ – after stratifying on age, the difference between drugs 2 and 3 became more significant ($p = 0.03$).

32.6.5 Relative Survival

Ederer et al. (1961) proposed two methods of describing the relative survival of a group of cancer patients to that of a standard population. The idea is to divide the observed survival (as estimated by the Kaplan–Meier estimate) in one group of patients by their expected survival based on the experience of a reference group. A key question is how to estimate the expected survival. The Ederer I method simply uses the mean expected survival function: $S_\mathrm{e}(t) = \sum_{\{i=1,\dots,n\}} S_i(t)/n$ where $Si(t)$ is the probability of survival to time t for the ith individual based on the standard population. Here the mean is calculated based on all individuals in the cohort and does not take into account the different make-up of the cohort as individuals die (or are censored). By contrast, the Ederer II method takes the average hazard among those still at risk and converts that to a survival function (via the formula

$S(t) = \exp[-\Lambda(t)]$). The expected or conditional cumulative hazard function is estimated by

$$\Lambda_c(t) = \int_0^t \left\{ \sum Y_i(s)\lambda_i(s) \Big/ \sum Y_i(s) \right\} \mathrm{ds}, \tag{32.1}$$

where $\lambda_i(s)$ is the hazard function appropriate for the ith individual calculated from the reference population and $Y_i(s)$ is an indicator of whether the ith individual was at risk at time s. In the Ederer II method, the survival function does not correspond to the survival of an actual population but is a measure of the excess risk in a competing-risks model. Suppose the hazard in the study population is equal to the hazard in the reference population plus some excess, then the difference between the observed cumulative hazard and the conditional expected cumulative hazard will be equal to the cumulative excess hazard.

The difference between the two methods can be illustrated by reference to a group of cancer patients some of whom are aged 40–49 and some of whom are aged 80–89. In the Ederer I method, the 80- to 89-year-old patients will still have an impact on the relative survival after 10 years. In the Ederer II method, the older patients will have less and less impact on the relative survival as they die. So that, if after 6 years, there are no older patients still at risk, the hazards (both observed and expected) for 7–10 years will be based exclusively on the younger patients.

A slight variation on the Ederer II method was proposed by Hakulinen (1982). Instead of using an indicator of whether the ith individual was still at risk, he suggested using an estimate of whether the ith individual would still be at risk if she/he were subject to the survival of the reference population and the censoring from the study. That is, $Y_i(s)$ is replaced by $S_i(s)C_i(s)$ where $C_i(s)$ is an indicator of whether the ith individual would still have been under follow-up at time s. A problem with this approach is that it requires knowledge of the censoring times even of those individuals who are not censored. That is not a real problem if there is no loss to follow-up and the maximal follow-up time for an individual can be calculated, for instance, from his/her date of entry into the study. An advantage of Hakulinen's method over Ederer II is that the expected survival function can truly be interpreted as a survival function.

Application of these standard approaches is sometimes hampered by missing covariate data. However, statistical techniques to take this uncertainty into account are sometimes appropriate, and Nur et al. (2010) provide a tutorial.

Perme et al. (2012) have recently proposed another alternative. It uses an Ederer II relative survival-type comparison but with two adjustments. Firstly, the at-risk indicator $Y_i(s)$ is replaced by $Y_i(s)/S_i(s)$ in Eq. 32.1. Secondly, they also adjusted the Nelson–Aalen estimate for the group of cancer patients by weighting each event $N_i(s)$ (1 for event, 0 for no event) to be $N_i(s)/S_i(s)$, as well as weighting the at-risk indicator as just described. The idea is to recover the number of events and time at risk in the absence of death from other causes, under an assumption that the reweighting process is suitable. This idea is sometimes called inverse probability weighting. Roche et al. (2013) applied the method as well as the Ederer and the Hakulinen approaches to seven cancer types and compared estimates of 5-, 10-, and 15-year net survival estimates, finding little difference between them for 5-year survival but larger differences for 10- and 15-year survival.

32.7 Regression Analysis of Variables That Influence Survival

Regression models are important when there are several factors upon which the disease rates may depend. Chapter ▸Regression Methods for Epidemiological Analysis of this handbook gives a general introduction to regression models and also discusses some of those examined here. The first model reviewed in this section is Poisson regression. This approach is particularly useful when the mortality in the study population is of the same order of magnitude as that of a standard reference population. It is also sometimes used to make projections of future rates of disease. When considering a cohort of patients with a life-threatening disease, the background rate of mortality is often of little interest, or when studying events other than death, there may be no reference data available. Two sorts of regression model that are useful in these circumstances are also reviewed: accelerated failure time models that relate variables to the survival time and Cox and Aalen semi-parametric models that link variables to the hazard or rate function.

32.7.1 Poisson Regression

The basic observation for Poisson regression consists of numbers of events and numbers of person-years at risk together with a number of covariates. An observation could relate to a single individual (in which case the number of events would usually be 0 or 1), or it could be the sum of the observations from a number of individuals with common covariate values. The model assumes that the observed number of events follows a Poisson distribution with a particular mean given by the number of person-years at risk multiplied by the modeled disease rates. For mathematical simplicity, the usual regression model is multiplicative, that is, it is assumed that the logarithm of the hazard follows a linear regression model: $\ln(\lambda_i) = \sum_j x_{ij}\beta_j$. Such models can be fitted in many software packages. The observed numbers of events, O_i, are the values of the outcome variable, the x_{ij}'s are the covariate values, and the logarithms of person-years at risk p_i are given as offsets – their regression coefficient is forced to be one. In symbols, $E(O_i) = \exp\{\ln(p_i) + \sum_j x_{ij}\beta_j\}$.

In practice, it is often the case that the Poisson assumption is not appropriate. For whatever reasons, mortality and disease registry data often show greater variability than implied by the Poisson model. Various sophisticated solutions to the problem of "extra-Poisson variation" have been put forward, but simple solutions usually suffice. One approach assumes that the variance of O_i is directly proportional to (but not necessarily equal to) the expected value of O_i. Under that assumption, one can estimate the dispersion factor (the coefficient of proportionality), by dividing the Pearson chi-squared statistic for the "saturated model" (that is the model with the most terms in it or the one including all the explanatory variables) by the number of degrees of freedom (McCullagh and Nelder 1989). Inference proceeds by multiplying the model-based standard errors by the square-root of the dispersion factor. Although this quasi-likelihood approach often works well, it lacks a sound theoretical justification, and in practice, it may be difficult to define the saturated model. Another simple approach to dealing with extra-Poisson variation is to use the sandwich estimator of the variance instead of the model-based estimator (Huber 1967). The sandwich estimator, also known as the Huber estimator or even the robust estimator, is available in many statistical packages and provides valid asymptotic inference no matter what the true variance model (hence the name robust), provided the mean model is correct.

Example 32.3. Cervical cancer in England and Wales since 1950 (cont.)
The model used to explain changes in cervical cancer mortality rates during the second half of the twentieth century was

$$\begin{aligned}\ln(\text{rate}) = {} & f_1(\text{age}) + f_2(\text{cohort}) + f_{3a}(\text{year if age 20–49}) \\ & + f_{3b}(\text{year if age 50–69}) + f_{3c}(\text{year if age 70+})\ .\end{aligned}$$

Here the f's are functions that we estimate using cubic splines (f_1 and f_2) or step functions (f_{3a}, f_{3b}, and f_{3c}). Estimates of these functions are plotted in Fig. 32.5. Results summarizing the goodness of fit of this model compared to various sub-models are given in Table 32.7. The scaled deviance is the deviance divided by the dispersion factor from the full model. The dispersion factor is the Pearson chi-squared statistic divided by the degrees of freedom.

The model of a constant rate at all ages and at all times gives a hopelessly bad fit with a scaled deviance of over 40,000 on 671 degrees of freedom. Over 80% of the deviance can be explained simply by allowing the rates to vary with age, but the model still does not fit the data well. Nearly 80% of the remaining deviance (all but 1,194 of 5,789) can be explained by allowing the age-standardized rate to vary with birth cohort; and adding a main effect for period (year of death) leads only to a more modest improvement in the fit. The full model, with interactions between age and period, provides further improvements in the fit suggesting that the dominant factors are age and year of birth but that there have been significant changes over time and that these have not been the same across age groups. The accepted explanation for these age-specific period effects is the widespread use of screening in women age 20–64 from the mid-1980s (see Sasieni and Adams 1999). For instance, the relatively small period effect for ages 70+ can partly be explained by the fact that screening was only offered to women aged 20–64. This makes it unlikely to have had a large effect on mortality in older women, except possibly in recent times in which older women may have been screened several years earlier while still aged under 65 (cf. chapter ▶Screening of this handbook). Finally, note that even the "full" model (with 36 explanatory variables) still had considerable extra-Poisson variation – the dispersion factor was 1.74.

Poisson age–period–cohort models have been used to predict the future number of cases. This raises issues including whether it is better to project current trends or just use present rates and consideration of the most appropriate functional form of any projection. Møller et al. (2003) considered these and other aspects in an analysis of cancer projections against observations in Nordic countries. They found that projections were better than assuming current rates, and recommended certain types of projection. Their approach was used to make cancer incidence predictions in the UK to 2030 by Mistry et al. (2011). Some practical aspects implementing such models in the computer software Stata were described by Sasieni (2012).

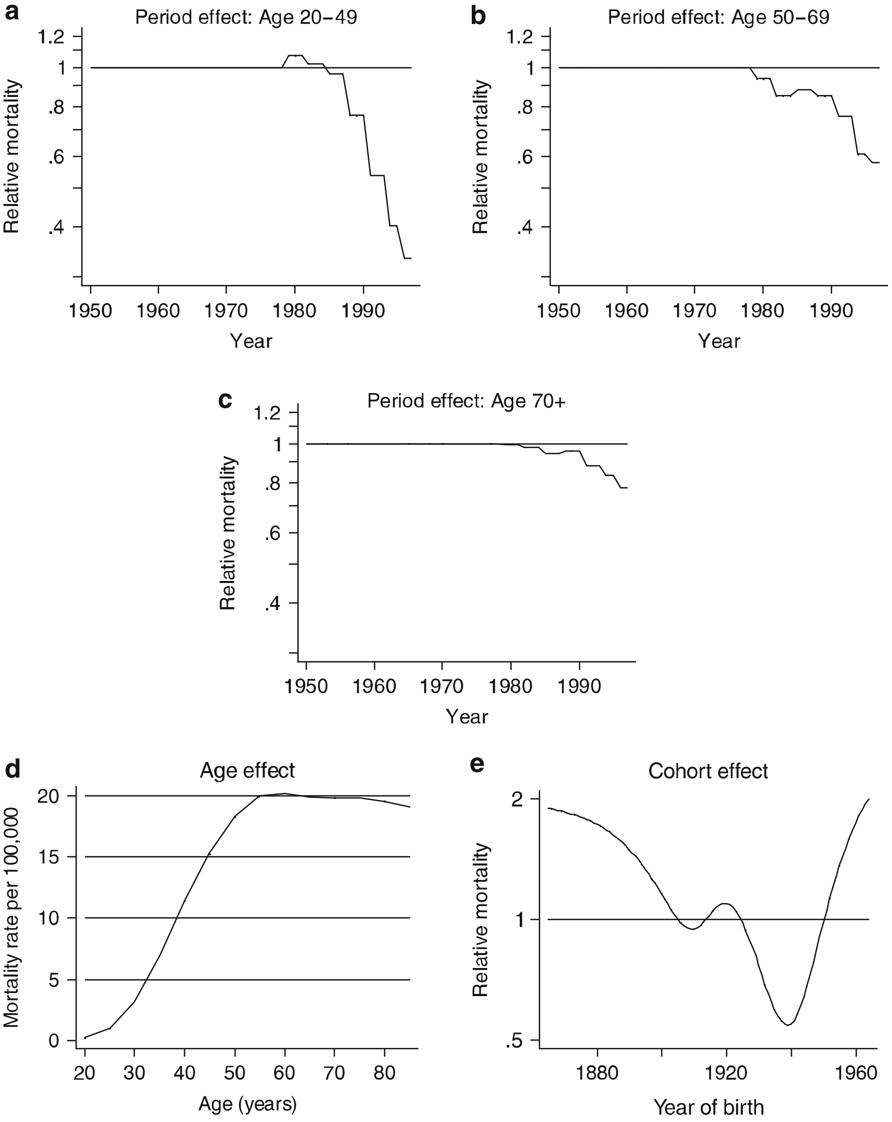


Fig. 32.5 Age, cohort, and age-specific period effects for cervical cancer mortality. The period effects are expressed as relative risks and are constrained to be 1.0 for 1950 to 1977. (**a**) Period effect for women aged 20–49; (**b**) Period effect for women aged 50–69; (**c**) Period effect for women aged 70+ (**d**) Age effect expressed as absolute rate for cohort born in 1905; (**e**) Cohort effect expressed as a relative risk

A common problem in using statistical software for Poisson regression is converting data on individuals including date of entry into the study, date of birth, date of exit and reason for exit (cause-specific death or censoring) into data suitable for Poisson regression: length of time at risk and number of events in

Table 32.7 Age–period–cohort modeling of cervical cancer mortality data from England and Wales

Model	Scaled deviance	Degrees of freedom	Chi-square/ df	Change in sc. deviance	Change in df
Null	41,288	671	95.66		
f_1(age)	5,789	665	14.71	35,499	6
f_1(age) + f_3(period)	2,629	658	7.02	3,160	7
f_1(age) + f_2(cohort)	1,194	656	3.10	4,595	9
f_1(age) + f_2(cohort) + f_3(period)	880	649	2.32	314	7
Full	648	635	1.74	232	14

Table 32.8 Example of how to code an individual for a Poisson model

Age group	Calendar year	Months at risk	Events
60–64	1986	7.5	0
60–64	1987	12	0
60–64	1988	10.5	0
65–69	1988	1.5	0
65–69	1989	12	0
65–69	1990	12	0
65–69	1991	12	0
65–69	1992	12	0
65–69	1993	10.5	0
70–74	1993	1.5	0
70–74	1994	12	0
70–74	1995	12	0
70–74	1996	12	0
70–74	1997	7.5	1

a number of risk groups. The situation can best be illustrated using a Lexis diagram (Keiding 1990; chapter ▸Descriptive Studies of this handbook). The horizontal axis is calendar time, the vertical axis is age, and individuals in the study can be represented by diagonal lines with slope 1. For instance, an individual could enter a study at age 62.5 in May 1986 and be followed until death in August 1997. Suppose population mortality rates are available in 5-year age bands (60–64, 65–69, 70–74, 75–79) for each calendar year. Then the individual's contributions to the various age groups and calendar years are as given in Table 32.8 (see also Fig. 32.6).

The conversion of individual data to age and calendar-year data is often referred to as person-years analysis and is fundamental to the analysis of large cohort studies. Programs exist in many software packages to facilitate the conversion of individual-level data (one record per person) to a format with a separate record for each individual in each risk stratum.

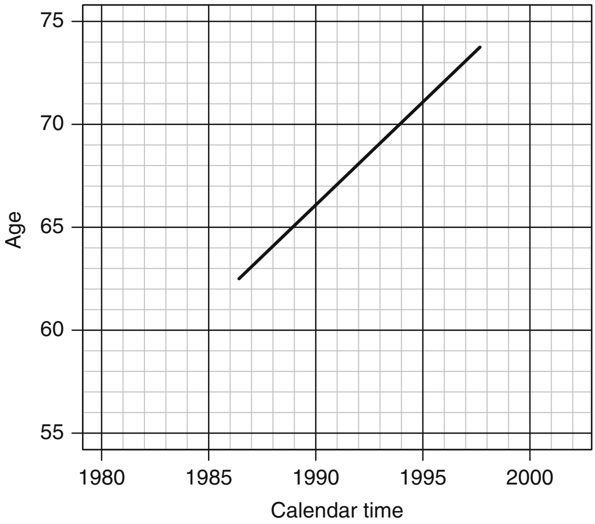


Fig. 32.6 Lexis diagram illustrating the experience of an individual, coded for a Poisson model in Table 32.8

32.7.2 Accelerated Failure Time Models

Most of the literature applying regression models to a function of the survival time assumes that the regression model is applied to the logarithm of the time. Such models are called "accelerated failure time models" because the effect of a covariate is to multiply the time scale on which events occur. It is as if the clock is made to run either slower or faster than usual. The simplest accelerated failure time model assumes that the survival times are exponential random variables, but other parametric models such as Weibull, gamma, and log-normal can also be used. For modeling all-cause adult mortality as a function of age, other distributions such as Gompertz or Makeham with hazards increasing exponentially with time are more appropriate. The book by Kalbfleisch and Prentice (2002) provides details of all these parametric distributions. All these regression models can be written as

$$\ln(T) = \beta_0 + \beta_1 Z_1 + \ldots + \beta_p Z_p + \varepsilon \, , \tag{32.2}$$

where $Z_1, \ldots, Z_p$ are covariates, $\beta_0, \ldots, \beta_p$ are regression parameters to be estimated, and ε is the random component distributed according to some model. (It should be noted that in this formulation, it is not ε but $\exp(\varepsilon)$ that has the named distribution.) A semi-parametric variant of this model uses the same formula (Eq. 32.2) but does not specify a parametric family for the distribution of ε. Parametric accelerated failure time models can be fitted using maximum likelihood estimation in a number of statistical packages. The semi-parametric model requires more advanced algorithms.

When using dummy covariates (i.e., Z_i is equal to 0 or 1), the regression parameters are easily interpreted: $\exp(\beta_i)$ is a multiplicative factor by which the "underlying" survival time is multiplied. If $\exp(\beta_i) = 3$, then on average individuals with $Z_i = 1$ take three times as long as individuals with $Z_i = 0$ to have the event of interest.

32.7.3 The Cox Model

Cox (1972) proposed modeling the conditional hazard as the product of an arbitrary baseline hazard $\lambda_0(t)$ and an exponential form that is linear in the covariates:

$$\lambda(t|Z_1, \ldots, Z_p) = \lambda_0(t) \exp\left(\beta_1 Z_1 + \ldots + \beta_p Z_p\right). \tag{32.3}$$

This model can be fitted quickly using most statistical software. We next comment on some aspects that might be important for applied epidemiology.

Relation to Other Models The restriction $\lambda_0(t) = \lambda_0$ leads to the exponential regression model (which is also an accelerated failure time model). Such a restriction will not generally be appropriate for problems in epidemiology where the underlying mortality rate is far from constant in time.

Taking $\lambda_0(t)$ to be the hazard from a reference population (i.e., discrete time) leads to the Poisson regression model (Breslow et al. 1983). Such modeling is particularly useful in epidemiology in which the goal may be to compare the survival of a study population to that of a reference population and to consider variables that affect the relative mortality. For this problem, Andersen et al. (1985) proposed a Cox-type model for the relative mortality replacing $\lambda_0(t)$ by the product of the individual-specific population mortality at time t, $\lambda_i^*(t)$, and an underlying relative mortality function, $\nu_0(t)$. They applied this model to patients with diabetes mellitus. Such a model may seem reasonable when studying a cohort at increased (or decreased) mortality from a number of causes, but the proportional excess mortality model (See 'Excess Risk' at the end of this section) may seem more appropriate when the cohort are at particular risk of certain causes of death and when that risk is unrelated to their underlying risk of death from other causes.

Censoring In the Cox model, there are three components to the data on each individual: the possibly censored failure time T; an indicator δ equal to 1 if T is a true failure time, 0 if it is censored; and $\vec{Z}$, the vector of explanatory variables. The model is flexible enough to incorporate explanatory variables that change value over the course of the study. The key censoring assumption is that the observation $(T = t, \delta = 0)$ tells us nothing more than that the true failure time is greater than t. When a study ends, some individuals will still be alive and will be censored. In such situations, it is necessary for survival to be independent of entry time for the above condition to be satisfied. Suppose for instance that early on after the introduction of a new type of surgery, only extremely sick patients are offered the new treatment but that once the hospital has 2 years experience in the technique, they offer it to relatively healthy patients too. If the results of all treated patients are studied 4 years after the introduction of the new technique, then patients censored in under 2 years may have been healthier at entry; thus if follow-up were continued for a further year, their survival in the third year might be better than that of patients entered in the first 2 years who survived 2 years. Thus the censoring time tells us something about the likely survival beyond the censoring time. In such circumstances, it is necessary to include "year of treatment" as an explanatory variable in the model.

We have discussed the effect of survival improving over calendar time when estimating survival rates, but when the duration of recruitment is short compared to the length of follow-up, such administrative censoring can usually be taken to be independent of survival. Other forms of censoring are more problematic. For instance, a patient who fails to attend a follow-up clinic might be too sick to get out of bed. So the fact that she/he was censored at t tells us more than simply that she/he was alive at t.

Relative Risk In the Cox model, $\lambda_0(t)$ is the hazard for an individual with $Z_i = 0$, $i = 1, \ldots, p$. Since all individuals with a $Z_i \neq 0$ are compared against $\lambda_0(t)$, it is called the baseline hazard. The model specifies that the hazard ratio between two individuals is constant over time. If one individual has covariates Z and other has covariates Z^*, then their hazard ratio is given by $\lambda_0(t|Z)/\lambda_0(t|Z^*) = \exp\{\beta(Z - Z^*)\}$, and it does not depend on the baseline hazard.

The relative risk is defined as the ratio of probabilities, that is, $\Pr(T < t|Z)/\Pr(T < t|Z^*)$. The hazard ratio will be a reasonable approximation of the relative risk provided that the event of interest is rare. To demonstrate, consider a hazard ratio of 2.0 and denote survival in the first individual to time t as S. Then the relative risk is $(1 - S^2)/(1 - S)$, or equivalently the relative risk of death in the other group will be $1 + S$ since $1 - S^2 = (1 + S)/(1 - S)$. Thus when the survival is 99% in one group, it will be 98.01% in the other group yielding a relative risk of 1.99, but when the survival in one group is 20%, it will be just 4% in the other group, yielding a relative risk of death of $0.96/0.80$ or 1.2.

Example 32.10. Testing a treatment for primary biliary cirrhosis (cont.)
Table 32.9 presents the results of fitting a Cox model to data from 216 patients with primary biliary cirrhosis in a clinical trial of azathioprine versus placebo (Christensen et al. 1985). The six variables were selected from an initial set of 25 partly using forward stepwise selection. An additional 32 patients were excluded because they had missing values of one or more of the six variables. The regression coefficients may be combined with their standard errors to obtain confidence intervals that rely on the asymptotic normality of the estimates. The positive coefficient associated with treatment implies that patients on the placebo (therapy = 1) had poorer prognosis than those on azathioprine (therapy = 0): the hazard of those on placebo is about 1.7 times greater than that of those on active treatment. Similarly, older patients had poorer prognosis. The hazard ratio associated with two patients aged 50 and 30 is $\exp[0.007\{\exp(3) - \exp(1)\}] = 1.13$. Notice, however, that the effect on survival is not fully described by the information in Table 32.9 because, without estimating the baseline hazard, one cannot translate the regression coefficients into effects on 5-year survival nor on median survival. Most statistical software for Cox regression will

Table 32.9 Cox model fitted to data from a clinical trial comparing the effects of azathioprine and placebo on the survival of 216 patients with primary biliary cirrhosis

Variable	Coding	$\hat{\beta}$	se($\hat{\beta}$)	exp($\hat{\beta}$)
Age	exp[(age in years − 20)/10]	0.007	0.0016	1.0
Albumin	serum value in gram per liter	−0.05	0.018	0.95
Bilirubin	$\log_{10}$ (serum concentration in micromole per liter)	2.51	0.32	12.3
Cholestasis	0 = no central cholestasis; 1 = yes	0.68	0.28	2.0
Cirrhosis	0 = no; 1 = yes	0.88	0.22	2.4
Therapy	0 = azathioprine; 1 = placebo	0.52	0.20	1.7

se standard error

also estimate the cumulative baseline hazard function $\Lambda_0(t)$ which is equal to the integral from 0 to t of $\lambda_0(u)$, and from this, one can calculate the estimated survival function for a given vector of covariates using the formula

$$\hat{S}(t|z) = \exp\{-\hat{\Lambda}_0(t)\exp(0.007\text{age} - 0.05\text{albumin} + 2.51\text{bilirubin} \\ + 0.68\text{cholestasi} + 0.88\text{cirrhosis} + 0.52\text{therapy})\}\ ,$$

where $\hat{\Lambda}_0$ is the estimate of Λ_0.

Stratified Model The basic model (32.3) has been generalized in various directions. A simple generalization is to permit different baseline hazard functions in each of a number of strata. The stratified Cox model assumes that, within each stratum, the proportional hazards assumption is justified and that the effect of the variable Z is the same in all strata:

$$\lambda(t|Z_1, \ldots, Z_p, \text{stratum } j) = \lambda_j(t)\exp\left(\beta_1 Z_1 + \ldots + \beta_p Z_p\right)\ . \qquad (32.4)$$

By incorporating constructed variables that are constant in some strata, the stratified model (32.4) can be used to model interactions between explanatory variables and strata. Suppose, for example, that one is stratifying by sex and including age as an explanatory variable. Let Z_1 = (age − 50) for men, = 0 for women; and let Z_2 = (age − 50) for women, = 0 for men. Then a model stratified on sex that includes Z_1, Z_2, and a treatment indicator Z_3 permits interactions between age and sex but assumes that the treatment acts proportionately on the hazards for any age–sex combination.

Link Functions Some epidemiologists like to use a different mathematical form of the explanatory variables in a Cox model, otherwise called a different link function. One version that is sometimes used is $\lambda(t|Z_1, \ldots, Z_p) = \lambda_0(t)(1 + \beta_1 Z_1 + \ldots + \beta_p Z_p)$. This is still a proportional hazards model, but the effect of different covariates is additive rather than multiplicative. For example if in model (32.3), being male infers a hazard ratio of 2 and being black infers a hazard ratio of 3, then being a black male infers a hazard ratio of 6 (= 2 × 3) relative to white females.

With the additive link, the hazard ratio for black males will be $1 + (2 - 1) + (3-1) = 4$. This form is useful when the variable is at a population level rather than an individual level, such as a percentage of the population exposed. For example, if there is a single term $Z_1(t)$ recording the proportion of individuals exposed to a factor at time t and exposure has a multiplicative effect on the individual hazard, then for a group of n-independent individuals, the appropriate hazard is $n\lambda_0(t)(1 + \beta_1 Z_1(t))$ where $\beta_1 \geq 0$ is the exposure effect. Duffy et al. (2007) considered this issue more generally.

Excess Risk When studying a cohort diagnosed with a particular disease or defined by exposure to a potential risk factor, it is often natural to model the excess mortality (or the excess cause-specific mortality). The reason is that we may assume that members of the cohort are subject to all the usual causes of death and that their "exposure" adds an independent route of death. For instance, patients with lymphoma might die from something unrelated or they might die of their lymphoma; the former will primarily be a function of age (mortality rises steeply with age), and the latter will largely depend on the time since their diagnosis. Gore et al. (1984) considered a variety of models for the analysis of survival in breast cancer patients. Sasieni (1996) showed how the Cox model could be applied to the excess mortality. The model was applied to nearly 1,000 patients with non-Hodgkins lymphoma. As one might have expected the effect of prognostic factors such as histology and stage were greater in the proportional excess model than in the usual Cox model (because these factors will not influence mortality from other causes). The model was also useful in showing that although there was still a deleterious effect of increasing age on the excess mortality, it was less than the effect estimated applying the Cox model to all-cause mortality.

32.7.4 Aalen Model

Aalen (1980, 1989, 1993) proposed an additive model for the conditional hazard function. His model is more general than the Cox model in that it has an unspecified function associated with each covariate. However (with two or more covariates), the Cox model is *not* a special case of the Aalen model. The Aalen model is that

$$\lambda(t|Z_1, \ldots, Z_p) = \lambda_0(t) + \lambda_1(t)Z_1 + \ldots + \lambda_p(t)Z_p\ , \tag{32.5}$$

where the functions $\lambda_0(t), \ldots, \lambda_p(t)$ are all unspecified and $\lambda_0(t)$ is the baseline hazard corresponding to an individual with $Z_i = 0, i = 1, \ldots, p$. It should be noted that if $\{Z_1, \ldots, Z_p\}$ is a set of dummy covariates for some factor, then the Aalen model is simply the non-parametric model allowing a different hazard function for each level of the factor. Biologically, the additive model may be interpreted in terms of excess risk. Such a model would be appropriate if each covariate contributed to a different route of death and these routes were independent of each other.

The fact that the Aalen model is so big (i.e., it has relative few constraints compared to a completely non-parametric model) is both to its advantage and its disadvantage. With no restriction on the relation between hazards over time, the model provides a description of the temporal influence of covariates. The Aalen model may be viewed as a one-step Taylor series approximation (i.e., a linear approximation) of an arbitrary $\lambda(t|Z_1, \ldots, Z_p)$ and will therefore provide a reasonable fit to any data provided the covariates have been centered and their effect is not too strong. For this reason, some suggest the use of the Aalen model as a diagnostic check in conjunction with the Cox model (Mau 1986; Henderson and Milner 1991). A disadvantage of the model is that the results cannot be presented in tabular form but require graphing of the estimated functions. It should also be noted that there is no restriction to prevent $\lambda_0(t) + \lambda_1(t)Z_1 + \ldots + \lambda_p(t)Z_p$ being negative (for some combination of covariate values), but a hazard function must be non-negative, so the model should not be applied to such covariate values.

The Aalen model has been used to model the excess mortality of cancer patients compared to that in the general population by Zahl (1996). He studied the long-term survival of men with colon cancer in Norway. A restriction of the Aalen model has been proposed by McKeague and Sasieni (1994). They suggested that some of the regression functions $\lambda_i(t)$ could be forced to be constant or a simple parametric function of time. Such a model allows parametric estimates of the constant additive effect of certain covariates. The special case $\lambda_0(t) + \beta_1 Z_1 + \ldots + \beta_p Z_p$, like the Cox model, has just one (baseline) hazard function to be estimated.

32.7.5 Using Regression Models to Adjust Survival Curves

When necessary, regression models can be used to adjust survival curves for confounding factors and thereby enable a more appropriate comparison to be made. To demonstrate, suppose we have a study designed to assess whether schizophrenic patients in a special residential environment were less likely to attempt suicide than patients cared for in the community. It might be impossible to set up a randomized study to carry out such a comparison. In this case, the statistical question is how to adjust the survival curve in patients cared for in the community so as to better reflect what one might have expected from the patients receiving residential care had they received usual care. One approach (Makuch 1982; Gail and Byar 1986) is to estimate the survival of each individual in the residential cohort using a model fitted to the standard cohort and to take the average of these survival estimates. For instance, if a Cox model is applied to the standard data with covariates Z_1 and Z_2 and yields estimates $\hat{\Lambda}(t)$ for the cumulative baseline hazard, $\hat{\beta}_1$ and $\hat{\beta}_2$, then the estimated survival curve for an individual with covariates z_{1i} and z_{2i} is

$$\hat{S}(t) = \exp\{-\hat{\Lambda}(t)\exp(\hat{\beta}_1 z_{1i} + \hat{\beta}_2 z_{2i})\} .$$

Example 32.7. Assessment of survival following liver transplantation (cont.)
Keiding et al. (1990) applied a Cox model to the English data with three covariates: ln(urea), ln(bilirubin), and an indicator for diuretic-responsive ascites. About 25% of the transplanted patients died within 2 months of transplant, but few patients died thereafter (the median follow-up was 6 months, the maximum was 3.5 years). This was quite similar to the "expected" survival based on the English data: the expected survival at 2 months was about 75% falling to about 60% by 6 months. The authors also compared the survival of their patients to that of primary biliary cirrhosis patients without transplant from an international trial of medical treatment. Once again, the expected survival was based on the results of fitting a Cox model. The expected survival based on the medically treated patients is similarly poor during the first 2 months but continues to decrease at a rapid rate so that the expected survival at 2 years is less than 20% compared with about 60% observed in the transplanted patients.

32.8 Further Extensions

In this necessarily brief overview of survival analysis, we focused attention on describing survival, comparing groups, and regression models. We end the chapter by commenting further on censoring, competing risks, and frailty.

32.8.1 Censoring and Truncation

Until now we have focused on *right-censoring*. The name comes from picturing a time line that runs left to right. The actual event time lies somewhere to the right of the censoring time, and information regarding what happens to the right of this time is censored. We have also discussed late-entry whereby some individuals do not join the study cohort at (their individual) time zero. Late entry is also sometimes called *left-truncation*. An example would arise in an epidemiological study of people living in a retirement community in which one was interested in factors influencing the age of death. In such a study, individuals only become eligible once they have moved to the retirement community. Some may move in at age 60, and others may not move there until they reach 80 (possibly following the death of a spouse). We observe how old individuals are when they join the community, but we do not know anything about individuals who die before they would have joined the community. If a woman joins at age 78, the fact that she lived to age 78 tells us nothing useful about events to the left of age 78 – had she died at age 76, she would not have been in the study. Both right-censoring and left-truncation can easily be accommodated by all the techniques that we have discussed thus far.

In epidemiology, one also has data that are subject to more extreme forms of censoring. The most extreme form arises from data on prevalence rates. In a cross-sectional study, one could survey whether individuals at different ages have or have not had some event in the past. For instance, one might be interested in age at menarche (and survey adolescent girls as to whether or not they have had their first period) or age at infection of a particular pathogen (and test for antibodies in blood). The assumption in such studies is that the event is not reversible and that the yes/no question will reduce the recall bias associated with asking people how long ago the event occurred. Such *current status* data are also referred to as (case I) *interval-censored* data. Here the age of the individual is the censoring time, and every individual is either left- or right-censored. Those who have not had the event are right-censored. But those who have are left-censored, that is, we know that their true event time lies somewhere to the left of the censoring time. Methods exist for analyzing prevalence data as reviewed by Keiding (1991). The classical problem is the non-parametric estimation of the survival function. The non-parametric maximum likelihood estimator is an extremely good estimator, and it can be estimated efficiently using the "pool adjacent violators algorithm" (Barlow et al. 1972). Testing for differences in survival based on interval-censored data was first discussed by Peto and Peto (1972). Several authors have considered regression models for prevalence data. Technically, fitting parametric regression models with prevalence data by maximum likelihood estimation presents no new problems (Odell et al. 1992). Semi-parametric models present more challenges (Huang 1996). The most widely studied are the Cox (proportional hazards) model and the proportional odds model. Essentially one can alternate between estimating the baseline hazard (or odds) for fixed values of the regression parameters and estimating the regression parameters for fixed value of the baseline function. Lin et al. (1998) proposed using the additive hazards model: $\lambda_0(t|Z) = \lambda_0(t) + \beta_1 Z_1(t) + \ldots + \beta_p Z_p(t)$ for current status data. They showed that the additive hazard model for current status data was equivalent to the usual Cox model for the event "observation made and observed to have died" with the covariate equal to minus the cumulative covariate from the additive hazard model (i.e., the integral from 0 to t of $-Z(s)$). The simple approach proposed by Lin et al. is not efficient and breaks down if the observation times are not independent of the covariates. Martinussen and Scheike (2002) showed how the additive hazard model could be fitted efficiently even in the presence of dependent observation times. Their method cannot simply be applied using standard Cox model software, but it is considerably simpler than efficient approaches to fitting the proportional hazards model to current status data. Groeneboom et al. (2010) proposed two estimators of the survivor function with interval-censoring, a so-called maximum smoothed likelihood estimator and a smoothed maximum likelihood estimator.

In other studies, individuals are screened or questioned periodically. In such studies, event times can be left- (the event happened before the first screen) or right-censored (the event happened after the last screen) or interval-censored (the event happened between two screens). This sort of data is sometimes called interval-censoring case 2 (Groeneboom and Wellner 1992) and sometimes *panel count* data. An added complication that is rarely considered is that the screening test may be less than 100% sensitive. In such circumstances, one should take into account the timing of all previous negative screens, not just the one immediately prior to the one that led to the detection of the event. For case 2 interval-censoring, the non-parametric maximum likelihood estimator of the survival function is harder to compute, but efficient algorithms exist (Groeneboom and Wellner 1992). Once again, testing has been considered (Zhang et al. 2001), and adaptation of the Cox model is possible

(Kooperberg and Clarkson 1997; Goetghebeur and Ryan 2000). A simple solution is to use conditional logistic regression to fit a proportional odds model (Rabinowitz et al. 2000). Rank estimation of a log-linear regression model has also been proposed (Li and Pu 2003). Practical issues such as how to deal with left-truncation in addition to interval-censoring and the effect of changes in disease incidence on the analysis are considered by Williamson et al. (2001).

A further complication arises when the initiating time is also possibly censored. For instance, one might be interested in the distribution of time from an infection to the development of symptomatic disease. It is likely that the time of infection will not be observed directly but will be interval-censored and the time of symptomatic disease may be right-censored. Such data are called *doubly-censored* (De Gruttola and Lagakos 1989).

Example 32.6. Blood transfusions and AIDS (cont.)
Each patient received blood over a period of time before HIV testing on donor blood was carried out, and so the time of infection was interval-censored. The time of infection was right-censored until symptoms were observed. Nearly one in six developed AIDS during follow-up, and a difference was observed between lightly and heavily treated patients.

Example 32.9. Skin cancer (cont.)
The time of metastasis was interval-censored and the time of death was right-censored. The authors used a Weibull model and looked at the effect of treatment, Breslow thickness, sex, site of the metastasis, and the time with stage II disease on survival post metastasis.

32.8.2 Competing Risks

Competing risks are nearly always present in survival analysis, but they may sometimes be dealt with by using an assumption about censoring or by using a comparative measure such as an excess hazard. It may also be possible to design a study with one or more risks removed. For example, one would not need to consider the competing risk of death from being run over by a bus for inhabitants of the Amazon jungle. Examples in epidemiology of when this may be useful are with a group of individuals that are relatively isolated and are therefore not exposed to some diseases or when a group is identified with genetic factors that protect against disease.

However, there are situations in which it is not possible to finesse the issue, and explicit modeling is required. A classic example is a three-state model in which patients start in remission and can either relapse or die while in remission (Fig. 32.7). One would certainly not wish to consider death as a source of independent censoring.

Whenever it is not possible to ignore the competing-risk problem, a common approach in epidemiology is to estimate the *cause-specific hazard* and from that the *cause-specific cumulative incidence*. Demographers and actuaries have traditionally called such quantities *crude*, meaning the joint function of times and causes; they contrast it with *net* for the marginals (integrating out the cause). An example of this setup is as follows.

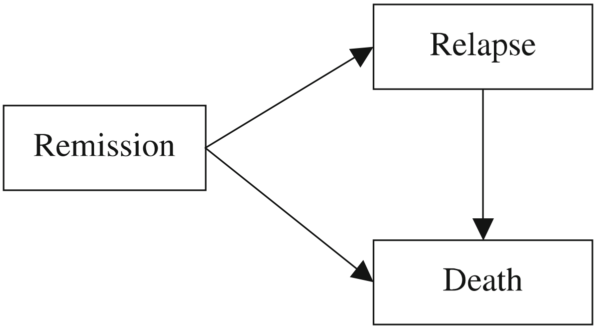


Fig. 32.7 Three-state model

Suppose individuals can die from one of K causes and that there is an additional process that (right) censors the data before they die. For example, individuals might be censored due to the end of follow-up in a clinical trial; we assume that the censoring is independent of mortality. The observed data on the ith individual consist of a survival time T_i and a cause indicator L_i with $L_i = 0$ if T_i is a censoring time. For cause $l \in \{0, 1, \ldots, K\}$, define the cause-specific hazard λ_l as one over Δ times the conditional probability of dying from cause l in the small interval between t and $t + \Delta$ conditional on surviving until time t: $\lambda_l = \Pr\{T < t + \Delta, L = l | T \geq t\}/\Delta$. The all-cause survival function S is defined in the usual way. The crude cumulative incidence function I_l for failures of type l is the probability that an individual will die from cause l by time t: $I_l(t) = \Pr\{T < t, L = l\}$. Let $\hat{S}$ denote the Kaplan–Meier estimator and d_{lj} and n_j the number of deaths of type l and the number at risk at time T_j, respectively. Note that $\hat{S}(T_{j-1})d_{lj}/n_j$ estimates probability of being alive at T_j times the conditional probability of dying from cause l between T_j and T_{j+1} given survival until T_j. Thus $I_l(t)$ can be estimated by $\sum_{\{j:T_j<t\}} \hat{S}(T_{j-1})d_{lj}/n_j$.

Example 32.15. Lung cancer trial

Lagakos (1978) tabulated and analyzed survival data from 194 patients with squamous cell carcinoma enrolled in a clinical trial. Of these, 83 had the event of local spread (cause 1), 44 had metastatic spread (cause 2), and 67 were censored (cause 0). An exponential model with constant cause-specific hazards was used, that is, where $\lambda_l(t) = \lambda_l$ for $l = 1, 2$ for all t. The main aim was to assess if three covariates were related to the survival: performance status (x_1, binary), treatment (x_2, binary), and age in years (x_3). They were included in the model via $\lambda_l = \exp(\beta_{l0} + \beta_{l1}x_1 + \beta_{l2}x_2 + \beta_{l3}x_3)$ for both $l = 1, 2$ causes, and the inferential problem was to assess whether $\beta_{lj} = 0$ for the $j = 1, 2, 3$ covariates. On the basis of the model fit, Lagakos found some evidence of a relationship between performance status and both causes but little for treatment or age. The data were reanalyzed in Crowder (2001), who first assessed whether the exponential model was justified. This was approached by considering a larger model in which the exponential form is a special case. He took $\lambda_l(t) = \phi_l \lambda_l^{\phi_l} t^{\phi_l - 1}$ where ϕ_1, ϕ_2 are additional parameters to be estimated, so that if $\phi_l = 1$, then the exponential model for cause l is obtained. Comparing the log-likelihood from this model (-35.21) with that from the exponential model (-45.57) yielded a log-likelihood ratio $\chi^2 = 2 \times (45.57 - 35.21) = 20.72$ on two degrees of freedom, so there was some evidence that the exponential model was not suitable ($p < 0.0001$). Crowder (2001) went on to consider further sub-models and tests of the covariates and ended up with similar conclusions to those in Lagakos (1978): the only important factor appeared to be performance status.

Further reading on competing risks is in textbooks from Crowder (2001, 2012).

32.8.3 Frailty

It is sometimes necessary to account for additional heterogeneity or lack of independence by allowing some individuals to be more "frail" than others. For instance, a standard life table assumes that all the variability within each cell can be explained by an exponential distribution, so that the hazard for each person i in a cell is the same λ for everyone. But, it might be more appropriate to suppose that the hazard for each person i is $u_i\lambda$, where $u_i \geq 0$ is the so-called frailty that has a distribution over the population. In other words, some individuals or groups are more likely to have the event than others. Allowing for frailty can make inference more robust but also changes the interpretation of observed trends. An example is where the rate of mortality is high at birth, then decreases during childhood and increases thereafter. A first interpretation might be that each infant faces a common hurdle to overcome in their first year of life. The alternative frailty interpretation is that some are more likely to die soon after birth than others, and once the frail babies have died, the healthy ones are still alive.

When the groups of individuals that might have the same frailty are specified in advance so that the hazard is $u_i\lambda_0(t)\exp(\beta_1 Z_1 + \ldots + \beta_p Z_p)$ for each individual in group i, where $\lambda_0(t)$ is an unknown baseline hazard function and $Z_1, \ldots, Z_p$ are known covariates, then a relatively popular extension of the Cox model to estimate the unknown parameters $\beta_1, \ldots, \beta_p$ allows a gamma frailty distribution with unit mean over the population (Clayton and Cuzick 1985). This model can be fitted in most statistical packages, which might explain part of its appeal. In Stata, it is fitted by the command `stcox` using the `shared` option to specify the groups. The gamma distribution is often justified because it is relatively flexible, but other frailty distributions that might be motivated from theoretical considerations have also been assessed, including the stable distribution. These are not currently available in Stata under a Cox model, for example, but when a parametric model is used instead then fitting the model is straightforward: the likelihood is written down, and a general-purpose optimization algorithm can be used to fit parameters.

Example 32.11. Infant mortality (cont.)
Moger and Aalen (2005) used a Weibull hazard and considered three possible parametric frailty distributions to account for common parents: (i) gamma, (ii) power variance function, and (iii) compound Poisson-gamma. All models had two parameters for the Weibull hazard. The first model had one parameter for the frailty distribution, so three in total; the second and third models had four and five parameters in total. The authors used standard routines in the statistical software S-PLUS to fit the models, based on derived likelihood functions. The log-likelihood value for model (i) was −4,865.0, for (ii) it was −4,861.9, and for (iii) it was −4,841.0. Model 3 fitted the data much better, for instance, the log-likelihood ratio

χ^2 for models 3 versus 2 was $2 \times (4{,}841.0 - 4{,}861.9) = 41.9$ on one degree of freedom, which is highly significant ($p < 0.0001$). A possible reason why the third model fitted the data best is that it allowed a proportion of the infants to be non-susceptible.

It might be of interest to estimate whether or not a frailty distribution improves a parametric survival model but without specifying the groups or a parametric frailty distribution. The non-parametric maximum likelihood estimate of the frailty distribution can be estimated using an expectation–maximization algorithm as described by Laird (1978) or by better algorithms mentioned in Crowder (2012).

Multivariate survival analysis is of the joint distribution of $K > 1$ survival times $t_1, \ldots, t_K$ for each unit. For the complete data setup, this is often approached via the multivariate normal distribution, but in survival analysis, one has to deal with censored data and so other techniques are called for, including the use of frailties. For instance, suppose there are two survival times of interest t_1 and t_2. Then one might model the hazard of these events for individual i as respectively $u_i\lambda_1$ and $u_i\lambda_2$, so that the survival times are said to be conditionally independent given u_i. The shared frailty $u_i \geq 0$ leads to positive dependence between the paired survival times. The gamma frailty Cox model mentioned above could be used for this setup. Another approach is to specify that the hazards are $u_{1i}\lambda_1$ and $u_{2i}\lambda_2$, where the frailties arise from a bivariate distribution. There are many more modeling possibilities, but applied work often uses parametric hazard models for λ_1 and λ_2, where a piecewise constant baseline hazard can be useful.

A wide variety of frailty models has been developed to account for aspects including familial relationships and sequences of lifetimes such as time between cancer diagnosis and relapse and then time from relapse to death or the time between skin melanomas. Some books that provide much more detail and focus on multivariate survival analysis are Hougaard (2000) and Crowder (2012); Therneau and Grambsch (2000) also have some material on frailty aspects.

32.9 Conclusions

Survival analysis is more closely associated with clinical medical research than with epidemiology, but there are a number of situations in which survival analysis is needed in epidemiological research. Some form of survival analysis will be required in any cohort study, and many of the more complicated designs used in modern epidemiology require quite sophisticated analytical techniques. In this chapter, we have presented a range of survival analysis tools covering a range of study designs. Before using these techniques, the reader is advised to consult an expert or to read a more complete text, but it is hoped that this chapter will help epidemiologists who come across survival analysis in the writing of others and those who think that they may need survival analysis in their own research.

The notions that events occur in time and that causes precede their effects are central to epidemiological research and the objective of understanding the causes of disease. Longitudinal studies are essential to epidemiology, and the complex evolution of risk factors, disease markers, and disease over time requires sophisticated statistical techniques.

References

Aalen OO (1980) A model for nonparametric regression analysis of counting processes. In: Klonecki W, Kozek A, Rosinski J (eds) Mathematical statistics and probability theory. Lecture notes in statistics, vol 2. Springer, New York, pp 1–25

Aalen OO (1989) A linear regression model for the analysis of life times. Stat Med 8:907–925

Aalen OO (1993) Further results on the non-parametric linear regression model in survival analysis. Stat Med 12:1569–1588

Andersen PK, Borch-Johnsen K, Deckert T, Green A, Hougaard P, Keiding N, Kreiner S (1985) A Cox regression model for the relative mortality and its application to diabetes mellitus survival data. Biometrics 41:921–932

Barlow RE, Bartholomew DJ, Bremner JM, Brunk HD (1972) Statistical inference under order restrictions. Wiley, Chichester

Borgan O, Langholz B (1993) Nonparametric estimation of relative mortality from nested case-control studies. Biometrics 49:593–602

Borgan O, Olsen EF (1999) The efficiency of simple and counter-matched nested case-control sampling. Scand J Stat 26(4):493–509

Brenner H (2002) Long-term survival rates of cancer patients achieved by the end of the 20th century: a period analysis. Lancet 360:1131–1135

Brenner H, Gefeller O (1996) An alternative method to monitor cancer patient survival. Cancer 78:2004–2010

Breslow NE, Day NE (1980) Statistical methods in cancer research. Volume I – The analysis of case-control studies. International Agency for Research on Cancer, Lyon. IARC Scientific Publications No. 32

Breslow NE, Day NE (1987) Statistical methods in cancer research. Volume II – The design and analysis of cohort studies. International Agency for Research on Cancer, Lyon. IARC Scientific Publications No. 82

Breslow NE, Lubin JH, Marek P, Langholz B (1983) Multiplicative models and cohort analysis. J Am Stat Assoc 78:1–12

Chen K, Lo S-H (1999) Case-cohort and case-control analysis with Cox's model. Biometrika 86:755–764

Chen Y-H (2002) Cox regression in cohort studies with validation sampling. J Roy Stat Soc B 64:51–62

Christensen E, Neuberger J, Crowe J, Altman DG, Popper H, Portmann B, Doniach D, Ranek L, Tygstrup N, Williams R (1985) Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis. Final results of an international trial. Gastroenterology 89:1084–1091

Clayton D, Cuzick J (1985). Multivariate generalizations of the proportional hazards model (with discussion). J Roy Stat Soc A 148:82–117

Cox DR (1972) Regression models and life tables (with discussion). J Roy Stat Soc B 34:187–220

Crowder MJ (2001) Classical competing risks. Chapman & Hall/CRC, Boca Raton

Crowder MJ (2012) Multivariate survival analysis and competing risks. Chapman & Hall/CRC, Boca Raton

Cuzick J (2008) Primary endpoints for randomised trials of cancer therapy. Lancet 371(9631):2156–2158

Cuzick J, Sasieni P, Evans S (1992) Ingested arsenic, keratoses, and bladder cancer. Am J Epidemiol 136:417–421

Cuzick J, Stewart H, Rutqvist L, Houghton J, Edwards R, Redmond C, Peto R, Baum M, Fisher B, Host H (1994) Cause-specific mortality in long-term survivors of breast cancer who participated in trials of radiotherapy. J Clin Oncol 12(3):447–453

Day NE (1976) A new measure of age standardized incidence, the cumulative rate. In: Waterhouse JAH, Muir CS, Correa P, Powell J (eds) Volume III – Cancer incidence in five continents. International Agency for Research on Cancer, Lyon. IARC Scientific Publications No. 15, pp 443–452

De Gruttola V, Lagakos SW (1989) Analysis of doubly-censored survival data, with application to AIDS. Biometrics 45:1–11

Dove-Edwin I, Sasieni P, Adams J, Thomas HJW (2005) Prevention of colorectal cancer by colonoscopic surveillance in individuals with a family history of colorectal cancer: 16 year, prospective, follow-up study. BMJ 331:1047

Duffy SW, Jonsson H, Agbaje OF, Pashayan N, Gabe R (2007) Avoiding bias from aggregate measures of exposure. J Epidemiol Community Health 61:461–463

Ederer F, Axtell LM, Cutler SJ (1961) The relative survival rate: a statistical methodology. Natl Cancer Inst Monogr 6:101–121

Gail MH, Byar DP (1986) Variance calculation for direct adjusted survival curves with application to testing for no treatment effect. Biom J 28:587–599

Galimberti S, Sasieni P, Valsecchi MG (2002) A weighted Kaplan–Meier estimator for matched data with application to the comparison of chemotherapy and bone-marrow transplant in leukaemia. Stat Med 21(24):3847–3864

Goetghebeur E, Ryan L (2000) Semiparametric regression analysis of interval-censored data. Biometrics 56:1139–1144

Gore SM, Pocock SJ, Kerr GR (1984) Regression models and non-proportional hazards in the analysis of breast cancer survival. Appl Stat 33:176–195

Greenwood M (1926) A report on the natural duration of cancer. Reports on Public Health and Medical Subjects, vol 33. H. M. Stationery Office, London, pp 1–26

Groeneboom P, Wellner J (1992) Information bounds and nonparametric maximum likelihood estimation. In: DMV Seminar, Band 19. Birkhauser, New York

Groeneboom P, Jongbloed G, Witte BI (2010) Maximum smoothed likelihood estimation and smoothed maximum likelihood estimation in the current status model. Ann Stat 38:352–387

Hakulinen T (1982) Cancer survival corrected for heterogeneity in patient withdrawal. Biometrics 38(4):933–942

Hall WJ, Wellner JA (1980) Confidence bands for a survival curve from censored data. Biometrika 67:133–143

Halley E (1693) An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the city of Breslaw; with an attempt to ascertain the price of annuities upon lives. Philos Trans R Soc Lond 17:596–610

Harwood CA, Mesher D, McGregor JM, Mitchell L, Leedham-Green M, Raftery M, Cerio R, Leigh IM, Sasieni P, Proby CM (2013) A surveillance model for skin cancer in organ transplant recipients: a 22-year prospective study in an ethnically diverse population. Am J Transplant. epub ahead of print. doi:10.1111/j.1600-6143.2012.04292.x

Henderson R, Milner A (1991) Aalen plots under proportional hazards. Appl Stat 40:401–409

Hosmer DW, Lemeshow S (1999) Applied survival analysis: regression modeling of time to event data. Wiley, New York

Hougaard P (2000) Analysis of multivariate survival data. Springer, New York

Huang J (1996) Efficient estimation for the Cox model with interval censoring. Ann Stat 24: 540–568

Huber PJ (1967) The behavior of maximum likelihood estimates under non-standard conditions. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol 1. University of California Press, Berkeley, pp 221–233

Kalbfleisch JD, Prentice RL (2002) The statistical analysis of failure time data, 2nd edn. Wiley, New York

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 58:457–481

Keiding N (1987) The method of expected number of deaths, 1786–1886–1986. Int Stat Rev 55: 1–20

Keiding N (1990) Statistical inference in the Lexis diagram. Philos Trans R Soc Lond A 332: 487–509

Keiding N (1991) Age-specific incidence and prevalence: a statistical perspective (with discussion). J R Stat Soc A 154:371–412

Keiding S, Ericzon BG, Eriksson S, Flatmark A, Höckerstedt K, Isoniemi H, Karlberg I, Keiding N, Olsson R, Samela K, Schrumpf E, Söderman C (1990) Survival after liver transplantation of patients with primary biliary cirrhosis in the Nordic countries. Comparison with expected survival in another series of transplantations and in an international trial of medical treatment. Scand J Gastroenterol 25:11–18

Kooperberg C, Clarkson DB (1997) Hazard regression with interval-censored data. Biometrics 53:1485–1494

Lagakos SW (1978) A covariate model for partially censored data subject to competing causes of failure. Appl Stat 27:235–241

Laird N (1978) Nonparametric maximum likelihood estimation of a mixing distribution. J Am Stat Assoc 73:805–811

Langholz B, Borgan O (1995) Counter matching: a stratified nested case-control sampling method. Biometrika 82:69–79

Langholz B, Goldstein L (1996) Risk set sampling in epidemiologic cohort studies. Stat Sci 11: 35–53

Lawrence G, Wallis M, Allgood P, Nagtegaal I, Warwick J, Cafferty F, Houssami N, Kearins O, Tappenden N, O'Sullivan E, Duffy S (2009) Population estimates of survival in women with screen-detected and symptomatic breast cancer taking account of lead time and length bias. Breast Cancer Res Treat 116:179–185

Leung KM, Elashoff RM (1996) A three-state disease model with intervalcensored data: Estimation and applications to AIDS and cancer. Lifetime Data Anal 2:175–194

Li L, Pu Z (2003) Rank estimation of log-linear regression with interval-censored data. Lifetime Data Anal 9:57–70

Lin DY, Oakes D, Ying Z (1998) Additive hazards regression with current status data. Biometrika 85:289–298

Makuch RW (1982) Adjusted survival curve estimation using covariates. J Chronic Dis 35: 437–443

Mantel N (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 50:163–170

Martinussen T, Scheike TH (2002) Efficient estimation in additive hazards regression with current status data. Biometrika 89:649–658

Marubini E, Valsecchi MG (1995) Analysing survival data from clinical trials and observational studies. Wiley, Chichester

Mau J (1986) On a graphical method for the detection of time-dependent effects of covariates in survival data. Appl Stat 35:245–255

McCullagh P, Nelder JA (1989) Generalised linear models, 2nd edn. Chapman & Hall, London

McKeague IW, Sasieni PD (1994) A partly parametric additive risk model. Biometrika 81:501–514

Mistry M, Parkin DM, Ahmad AS, Sasieni PD (2011) Cancer incidence in the United Kingdom: projections to the year 2030. Br J Cancer 105:1795–1803

Møller B, Fekjær H, Hakulinen T, Sigvaldason H, Storm HH, Talbäck M, Haldorsen T (2003) Prediction of cancer incidence in the Nordic countries: empirical comparison of different approaches. Stat Med 22:2751–2766

Moger TA, Aalen OO (2005) A distribution for multivariate frailty based on the compound Poisson distribution with random scale. Lifetime Data Anal 11:41–59

Nur U, Shack LG, Rachet B, Carpenter JR, Coleman MP (2010) Modelling relative survival in the presence of incomplete data: a tutorial. Int J Epidemiol 39:118–128

Odell PM, Anderson KM, D'Agostino RB (1992) Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. Biometrics 48:951–991

Perme MP, Stare J, Estève J (2012) On estimation in relative survival. Biometrics 68:113–120

Peto R, Peto J (1972) Asymptotically efficient rank invariant test procedures (with discussion). J R Stat Soc A 135:185–206

Prentice RL (1986a) On the design of synthetic case-control studies. Biometrics 42:301–310

Prentice RL (1986b) A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 73:1–11

Rabinowitz D, Betensky RA, Tsiatis AA (2000) Using conditional logistic regression to fit proportional odds models to interval-censored data. Biometrics 56:511–518

Roche L, Danieli C, Belot A, Grosclaude P, Bouvier AM, Velten M, Iwaz J, Remontet L, Bossard N (2013) Cancer net survival on registry data: use of the new unbiased Pohar-Perme estimator and magnitude of the bias with the classical methods. Int J Cancer 132:2359–2369

Sasieni PD (1996) Proportional excess hazards. Biometrika 83:127–141

Sasieni PD (2003) On the expected number of cancer deaths during follow-up of an initially cancer-free cohort. Epidemiology 14:108–110

Sasieni PD (2012) Age-period-cohort models in stata. Stata J 12:45–60

Sasieni PD, Adams J (1999) Standardized lifetime risk. Am J Epidemiol 149:869–875

Sasieni PD, Adams J (2000) Analysis of cervical cancer mortality and incidence data from England and Wales: evidence of a beneficial effect of screening. J R Stat Soc A 163:191–209

Sasieni PD, Adams J, Cuzick J (2002) Avoidance of premature death: a new definition for the proportion cured. J Cancer Epidemiol Prev 7:165–171

Sasieni PD, Shelton J, Ormiston-Smith N, Thomson CS, Silcocks PB (2011) What is the lifetime risk of developing cancer?: The effect of adjusting for multiple primaries. Br J Cancer 105: 460–465

Self SG, Prentice RL (1988) Asymptotic distribution theory and efficiency results for case-cohort studies. Ann Stat 11:804–812

Selvin S (2008) Survival analysis for epidemiologic and medical research. Cambridge University Press, Cambridge

Suissa S (2008) Immortal time bias in pharmacoepidemiology. Am J Epidemiol 167:492–499

Therneau TM, Grambsch PM (2000) Modeling survival data: extending the Cox model. Springer, New York

Thomas DC (1977) Addendum to: "Methods of cohort analysis: appraisal by application to asbestos mining", by FDK Liddell, JC McDonald and DC Thomas. J R Stat Soc A 140:469–491

Williamson JM, Satten GA, Hanson JA, Weinstock H, Datta S (2001) Analysis of dynamic cohort data. Am J Epidemiol 154:366–372

Winnett A, Sasieni P (2002) Adjusted Nelson-Aalen estimates with retrospective matching. J Am Stat Assoc 97:245–256

Zahl P-H (1996) A linear non-parametric regression model for the excess intensity. Scand J Stat 23:353–364

Zhang Y, Liu W, Zhan Y (2001) A nonparametric two-sample test of the failure function with interval censoring case 2. Biometrika 88:677–686

Measurement Error 33

Jeffrey S. Buzas, Leonard A. Stefanski, and Tor D. Tosteson

Contents

J.S. Buzas (✉)
Department of Mathematics and Statistics, University of Vermont, Burlington, VT, USA

L.A. Stefanski
Department of Statistics, North Carolina State University, Carolina, NC, USA

T.D. Tosteson
Dartmouth College, Lebanon, NH, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_19,

33.1 Introduction

Factors contributing to the presence or absence of disease are not always easily determined or accurately measured. Consequently, epidemiologists are often faced with the task of inferring disease patterns using noisy or indirect measurements of risk factors or covariates. Problems of measurement arise for a number of reasons, including reliance on self-reported information, the use of records of suspect quality, intrinsic biological variability, sampling variability, and laboratory analysis error. Although the reasons for imprecise measurement are diverse, the inference problems they create share in common the structure that statistical models must be fit to data formulated in terms of well-defined but unobservable variables X, using information on measurements W that are less than perfectly correlated with X. Problems of this nature are called measurement error problems, and the statistical models and methods for analyzing such data are called measurement error models.

This chapter focuses on statistical issues related to the problems of fitting models relating a disease response variable Y to true predictors X and error-free predictors Z, given values of measurements W, in addition to Y and Z. Although disease status may also be subject to measurement error, attention is limited to measurement error in predictor variables. We further restrict attention to measurement error in continuous predictor variables. Categorical predictors are not immune from problems of ascertainment, but misclassification is a particular form of measurement error. Consequently, misclassification error is generally studied separately from measurement error; see chapter ▶Misclassification in this handbook.

A case-control study exhibiting measurement error was described in Karagas et al. (2002) and is briefly mentioned here to exemplify the notation. The purpose of the study was to assess the risk of bladder and two forms of non-melanoma skin cancer (Y's) to "true" arsenic exposure (X), adjusting for patient age (Z). True arsenic exposure was measured imprecisely through concentrations of arsenic in toenails (W).

This chapter is organized in three main sections. Section 33.2 defines basic concepts and models of measurement error and outlines the effects of ignoring measurement error on the results of standard statistical analyses. An important aspect of most measurement error problems is the inability to estimate parameters of interest given only the information contained in a sample of (Y, Z, W) values. Some features of the joint distribution of (Y, Z, X, W) must be known or estimated in order to estimate parameters of interest. Thus, additional data, depending on the type of error model, must often be collected. Consequently, it is important to include measurement error considerations when planning a study, both to enable application of a measurement error analysis of the data and to ensure validity of conclusions. Planning studies in the presence of measurement error is the topic of Sect. 33.3. Methods for the analysis of data measured with error differ according to the nature of the measurement error, the additional parameter-identifying information that is available, and the strength of the modeling assumptions appropriate for a particular problem. Section 33.4 describes a number of common approaches to the analysis of data measured with error, including simple, generally applicable bias-adjustment approaches, conditional likelihood, and full likelihood approaches.

This chapter is intended as an introduction to the topic. There are several books providing in-depth coverage of measurement error models. Fuller (1987) studies linear measurement error models. Carroll et al. (2006) provide detailed coverage of non-linear models as well as density estimation. Buonaccorsi (2010) describes additional methods for non-linear models and other topics including time series and misclassification. Gustafson (2004) covers measurement error from a Bayesian perspective with a focus on misclassification.

Other review articles addressing measurement error in epidemiology include Thürigen et al. (2000), Carroll (1998), Thomas et al. (1993) and Armstrong et al. (1992). Prior to the book by Fuller (1987), the literature on measurement error models was largely concerned with linear measurement error models and went under the name *errors-in-variables*. Chapter ▸Regression Methods for Epidemiological Analysis of this handbook presents additional topics in regression modeling.

33.2 Measurement Error and its Effects

This section presents the basic concepts and definitions used in the literature on non-linear measurement error models. The important distinction between differential and non-differential error is discussed first and is followed by a description of two important models for measurement error. The major effects of measurement error are described and illustrated in terms of multivariate normal regression models.

33.2.1 Differential and Non-Differential Error and Surrogate Variables

The error in W as a measurement of X is *non-differential* if the conditional distribution of Y given (Z, X, W) is the same as that of Y given (Z, X), that is, $f_{Y|ZXW} = f_{Y|ZX}$. When $f_{Y|ZXW} \neq f_{Y|ZX}$, the error is *differential*. The key feature of a non-differential measurement is that it contains no information for predicting Y in addition to the information already contained in Z and X. When $f_{Y|ZXW} = f_{Y|ZX}$, W is said to be a *surrogate* for X.

Many statistical methods in the literature on measurement error modeling are based on the assumption that W is a surrogate. It is important to understand this concept and to recognize when it is or is not an appropriate assumption. Non-differential error is plausible in many cases, but there are situations where it should not be assumed without careful consideration.

If measurement error is due solely to instrument or laboratory analysis error, then it can often be argued that the error is non-differential. However, in epidemiological applications, measurement error commonly has multiple sources, and instrument and laboratory analysis error are often minor components of the total measurement error. In these cases, it is not always clear whether measurement error is non-differential.

The potential for non-differential error is greater in case-control studies because covariate information ascertainment and exposure measurement follow disease response determination. In such studies, selective recall, or a tendency for cases to overestimate exposure, can induce dependencies between the response and the true exposure even after conditioning on true exposure.

A useful exercise for thinking about the plausibility of the assumption that W is a surrogate is to consider whether W would have been measured (or included in a regression model) had X been available. For example, suppose that the natural predictor X is defined as the temporal or spatial average value of a time-varying risk factor or spatially varying exposure (e.g., blood pressure, cholesterol, lead exposure, particulate matter exposure) and the observed W is a measurement at a single point in time or space. In such cases, it might be convincingly argued that the single measurement contributes little or no information in addition to that contained in the long-term average.

However, this line of reasoning is not foolproof. The surrogate status of W can depend on the particular model being fit to the data. For example, consider models where Z has two components; $Z = (Z_1, Z_2)$. It is possible to have $f_{Y|Z_1Z_2XW} = f_{Y|Z_1Z_2X}$ and $f_{Y|Z_1XW} \neq f_{Y|Z_1X}$. Thus, W is a surrogate in the full model including Z_1 and Z_2 but not in the reduced model including only Z_1. In other words, whether a variable is a surrogate or not depends on other variables in the model. A simple example illustrates this feature. Let $X \sim N(\mu_x, \sigma_x^2)$. Assume that $\epsilon_1, \epsilon_2, U_1$, and U_2 are mean zero normal random variables such that $X, \epsilon_1, \epsilon_2, U_1, U_2$ are mutually independent. Let $Z = X + \epsilon_1 + U_1$, $Y = \beta_1 + \beta_z Z + \beta_z X + \epsilon_2$ and $W = X + \epsilon_1 + U_2$. Then $E(Y|X) \neq E(Y|X, W)$ but $E(Y|Z, X, W) = E(Y|Z, X)$. The essential feature of this example is that the measurement error $W - X$ is correlated with the covariate Z. The presence or absence of Z in the model determines whether W is a surrogate or not. Such situations have the potential of arising in applications. For example, consider air pollution health effects studies. Suppose that X is the spatial average value of an air pollutant, W is the value measured at a single location, the components of Z include meteorological variables, and Y is a spatially aggregated measure of morbidity or mortality (all variables recorded daily, with X, W, and Z suitably lagged). If weather conditions influence both health and the measurement process (e.g., by influencing the spatial distribution of the pollutant), then it is possible that W would be a surrogate only for the full model containing Z.

With non-differential measurement error, it is possible to estimate parameters in the model relating the response to the true predictor using the measured predictor only, with minimal additional information on the error distribution, that is, it is not necessary to observe the true predictor. However, this is not generally possible with differential measurement error. In this case, it is necessary to have a validation subsample in which both the measured value and the true value are recorded. Data requirements are discussed more fully in Sect. 33.3. Much of the literature on measurement error models deals with the non-differential error, and hence that is the focus of this chapter. Problems with differential error are often better analyzed via techniques for missing data.

33.2.2 Error Models

The number of ways a surrogate W and predictor X can be related are countless. However, in practice, it is often possible to reduce most problems to one of two simple error structures. For understanding the effects of measurement error and the statistical methods for analyzing data measured with error, an understanding of the two simple error structures is generally sufficient.

33.2.2.1 Classical Error Model

The standard statistical model for the case in which W is a measurement of X in the usual sense is $W = X + U$, where U has mean zero and is independent of X. As explained in the preceding section, whether W is a surrogate or not depends on more than just the joint distribution of X and W. However, in the sometimes plausible case that the error U is independent of all other variables in a model, then it is non-differential and W is a surrogate. This is often called the classical error model. More precisely, it is an independent, unbiased, additive measurement error model. Because $E(W \mid X) = X$, W is said to be unbiased measurement of X.

Not all measuring methods produce unbiased measurements. However, it is often possible to calibrate a biased measurement resulting in an unbiased measurement. Error calibration is discussed later in greater detail.

33.2.2.2 Berkson Error Model

For the case of Berkson error, X varies around W and the accepted statistical model is $X = W + U$, where U has mean zero and is independent of W. For this model, $E(X \mid W) = W$, and W is called an unbiased Berkson predictor of X, or simply an unbiased predictor of X. The terminology results from the fact that the best squared-error predictor of X given W is $E(X \mid W) = W$.

Berkson (1950) describes a measurement error model which is superficially similar to the classical error model, but with very different statistical properties. He describes the error model for experimental situations in which the observed variable was controlled, hence the alternative name *controlled variable model*, and the error-free variable, X, varied around W. For example, suppose that an experimental design called for curing a material in a kiln at a specified temperature W, determined by thermostat setting. Although the thermostat is set to W, the actual temperature in the kiln, X, often varies randomly from W due to less-than-perfect thermostat control. For a properly calibrated thermostat, a reasonable assumption is that $E(X \mid W) = W$, which is the salient feature of a Berkson measurement (compare to an unbiased measurement for which $E(W \mid X) = X$).

Apart from experimental situations, in which W is truly a controlled variable, the unbiased Berkson error model seldom arises as a consequence of sampling design or direct measurement. However, like the classical error model, it is possible to calibrate a biased surrogate so that the calibrated measurement satisfies the assumptions of the Berkson error model.

33.2.2.3 Reduction to Unbiased Error Model

The utility of the classical and Berkson error structures is due to the fact that many error structures can be transformed to one or the other. Suppose that W^* is a surrogate for X. For the case that a linear model for the dependence of W^* on X is reasonable, that is, $W^* = \gamma_1 + \gamma_x X + U^*$, where U^* is independent of X, the transformed variable $W = (W^* - \gamma_1)/\gamma_x$ satisfies the classical error model $W = X + U$, where $U = U^*/\gamma_x$. In other words, W^* can be transformed into an independent, unbiased, additive measurement.

Alternatively, for the transformation $W = E(X \mid W^*)$, it follows that $X = W + U$, where $U = X - E(X \mid W^*)$ is uncorrelated with W. Thus, apart from the distinction between independence and zero correlation of the error U, any surrogate W^* can be transformed to an unbiased additive Berkson error structure.

Both types of calibration are useful. The transformation that maps an uncalibrated surrogate W^* into a classical error model is called *error calibration*. The transformation that maps W^* into a Berkson error model is called *regression calibration* (Carroll et al. 2006); see Tosteson et al. (1989) for an interesting application of regression calibration.

In theory, calibration reduces an arbitrary surrogate to a classical error measurement or a Berkson error measurement, explaining the attention given to these two unbiased error models. In practice, things are not so simple. Seldom are the parameters in the regression of W on X (error calibration) or in the regression of X on W (regression calibration) known, and these parameters have to be estimated, which is generally possible only if supplementary data are available for doing so. In these cases, there is yet another source of variability introduced by the estimation of the parameters in the chosen calibration function. This is estimator variability and should be accounted for in the estimation of standard errors of the estimators calculated from the calibrated data.

33.2.3 Measurement Error in the Normal Linear Model

We now consider the effects of measurement error in a simple linear regression model with normal variation. This model has limited use in epidemiology, but it is one of the few models in which the effects of measurement error can be explicitly derived and explained. Measurement error affects relative risk coefficients in much the same way as regression coefficients so that the insights gained from this simple model carry over to more useful epidemiological models.

Consider the multivariate normal formulation of the simple linear regression model,

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} \beta_1 + \beta_x \mu_x \\ \mu_x \end{pmatrix}, \begin{pmatrix} \beta_x^2 \sigma_x^2 + \sigma_\epsilon^2 & \beta_x \sigma_x^2 \\ \beta_x \sigma_x^2 & \sigma_x^2 \end{pmatrix} \right\}. \tag{33.1}$$

If, as is assumed here, the substitute variable W is jointly normally distributed with (Y, X), then in the absence of additional assumptions on the relationship between W and (Y, X), the multivariate normal model for $(Y,\ X, W)$ is

$$\begin{pmatrix} Y \\ X \\ W \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} \beta_1 + \beta_x \mu_x \\ \mu_x \\ \mu_w \end{pmatrix}, \begin{pmatrix} \beta_x^2 \sigma_x^2 + \sigma_\epsilon^2 & \beta_x \sigma_x^2 & \beta_x \sigma_{xw} + \sigma_{\epsilon w} \\ \beta_x \sigma_x^2 & \sigma_x^2 & \sigma_{xw} \\ \beta_x \sigma_{xw} + \sigma_{\epsilon w} & \sigma_{xw} & \sigma_w^2 \end{pmatrix} \right\}, \tag{33.2}$$

where $\sigma_{xw} = \mathrm{Cov}(X, W)$ and $\sigma_{\epsilon w} = \mathrm{Cov}(\epsilon, W)$. In measurement error modeling, the available data consist of observations (Y, W) so that the relevant sampling model is the marginal distribution of $(Y,\ W)$.

We now describe biases that arise from the so-called *naive* analysis of the data, that is, the analysis of the observed data using the usual methods for error-free data. In this case, the naive analysis is least squares analysis of $\{(W_i, Y_i),\ \ i = 1, \ldots, n\}$ so that the naive analysis results in unbiased estimates of the parameters in the regression model for Y on W, or what we refer to as the *naive model*. Naive-model parameters are given in Table 33.1 for some particular error models.

33.2.3.1 Differential Error

For the case of a general measurement with possibly differential error, the naive estimator of slope is an unbiased estimator of $(\beta_x \sigma_{xw} + \sigma_{\epsilon w})/\sigma_w^2$ rather than β_x. Depending on the covariances between ϵ and W, and X and W, and the variance of W, the naive-model slope could be less than or greater than β_x, so that no general conclusions about bias are possible. Similarly, the residual variance of the naive regression could be either greater or less than the true model residual variance. It follows that for a general measurement W, the coefficient of determination for the naive analysis could be greater or less than for the true model. These results indicate the futility of trying to make generalizations about the effects of using a general measurement for X in a naive analysis.

33.2.3.2 Surrogate

For the multivariate normal model with $0 < \rho_{xw}^2 < 1$, W is a surrogate if and only if $\sigma_{\epsilon w} = 0$. With an arbitrary surrogate measurement, the naive estimator of slope unbiasedly estimates $\beta_x \sigma_{xw}/\sigma_w^2$. Depending on the covariance between X and W and the variance of W, the naive-model slope could be less or greater than β_x, so that again no general statements about bias in the regression parameters are possible. For an uncalibrated measurement, $E(W|X) = \gamma_0 + \gamma_x X$, $\sigma_{xw} = \mathrm{cov}(X, W) = \gamma_x \sigma_x^2$ and $\mathrm{Var}(X) = \gamma_x^2 \sigma_x^2 + \sigma_u^2$. In this case, the relative bias, $\sigma_{xw}/\sigma_w^2 = \gamma_x \sigma_x^2/(\gamma_x^2 \sigma_x^2 + \sigma_u^2)$, is bounded in absolute value by $1/|\gamma_x|$. For an uncalibrated Berkson measurement, $E(X|W) = \alpha_1 + \alpha_w W$, $\sigma_{xw} = \alpha_w \sigma_w^2$, and the relative bias is α_w. When W is a surrogate, the residual variance from the naive analysis is never less than the true model residual variance and is strictly greater except in the extreme case that X and W are perfectly correlated, $\rho_{xw}^2 = 1$. It follows that for an arbitrary surrogate, the coefficient of determination for the naive model is always less than or equal to that for the true model. The use of a surrogate always entails a loss of predictive power. The naive-model slope indicates that in order to recover β_x from an analysis of the observed data, only σ_{xw} would have to be known. A *validation study* in which bivariate observations (X, W) were obtained would provide the necessary information for estimating σ_{xw}.

33.2.3.3 Classical Error

If the surrogate, W, is an unbiased measurement, $E(W \mid X) = X$, and the classical error model holds, then $\mu_w = \mu_x$, $\sigma_{xw} = \sigma_x^2$, and $\sigma_w^2 = \sigma_x^2 + \sigma_u^2$. In this case, the naive slope estimator unbiasedly estimates $\beta_x \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$. For this case, the sign ($\pm$) of $\beta_x \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$ is always the same as the sign of β_x, and the inequality $\sigma_x^2/(\sigma_x^2 + \sigma_u^2)|\beta_x| \leq |\beta_x|$ shows that the naive estimator of slope is always biased toward 0. This type of bias is called *attenuation* or *attenuation toward the null*. The attenuation factor $\lambda = \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$ is called the *reliability ratio*, and its inverse is called the *linear correction for attenuation*. In this case, the coefficient of determination is also attenuated toward zero, and the term attenuation is often used to describe both attenuation in the slope coefficient and the attenuation in the coefficient of determination. *Regression dilution* has also been used in the epidemiology literature to describe attenuation (MacMahon et al. 1990; Hughes 1993). In order to recover β_x from an analysis of the observed data, it would be sufficient to know σ_u^2. Either replicate measurements or validation data provide information for estimating the measurement error variance σ_u^2.

33.2.3.4 Berkson Error

With W a surrogate, the Berkson error model is embedded in the multivariate normal model by imposing the condition $E(X \mid W) = W$. In this case, $\mu_x = \mu_w, \sigma_{xw} = \sigma_w^2$ and $\sigma_x^2 = \sigma_w^2 + \sigma_u^2$. When W and X satisfy the unbiased Berkson error model, $X = W + U$, the naive estimator of slope is an unbiased estimator of β_x, that is, there is no bias. Thus, there is no bias in the naive regression parameter estimators, but there is an increase in the residual variance and a corresponding decrease in the model coefficient of determination. Even though no bias is introduced, there is still a penalty incurred with the use of Berkson predictors. However, with respect to valid inference on regression coefficients, the linear model is robust to Berkson errors. The practical importance of this robustness property is limited because the unbiased Berkson error model seldom is appropriate without regression calibration except in certain experimental settings as described previously.

33.2.3.5 Discussion

Measurement error is generally associated with attenuation, and as the Table 33.1 shows, attenuation in the coefficient of determination occurs with any surrogate measurement. However, attenuation in the regression slope is, in general, specific only to the classical error model. The fact that measurement-error-induced bias

Table 33.1 Table entries are slopes and residual variances of the linear model relating Y to W for the cases in which W is a differential measurement, a surrogate, an unbiased classical error measurement, an unbiased Berkson predictor, and the case of no error ($W = X$)

Error model	Slope	Residual variance
Differential	$\beta_x\left(\frac{\sigma_{xw}}{\sigma_w^2}\right)+\left(\frac{\sigma_{\epsilon w}}{\sigma_w^2}\right)$	$\sigma_\epsilon^2+\beta_x^2\sigma_x^2-\frac{(\sigma_{xw}\beta_x+\sigma_{\epsilon w})^2}{\sigma_w^2}$
Surrogate	$\beta_x\left(\frac{\sigma_{xw}}{\sigma_w^2}\right)$	$\sigma_\epsilon^2+\beta_x^2\sigma_x^2(1-\rho_{xw}^2)$
Classical	$\beta_x\left(\frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}\right)$	$\sigma_\epsilon^2+\beta_x^2\sigma_x^2\left(\frac{\sigma_u^2}{\sigma_x^2+\sigma_u^2}\right)$
Berkson	β_x	$\sigma_\epsilon^2+\beta_x^2\sigma_x^2\left(\frac{\sigma_u^2}{\sigma_x^2}\right)$
No error	β_x	σ_ϵ^2

depends critically on the type of measurement error underlies the importance of correct identification of the measurement error in applications. Incorrect specification of the measurement error component of a model can create problems as great as those caused by ignoring measurement error.

The increase in residual variance associated with surrogate measurements (including classical and Berkson) gives rise not only to a decrease in predictive power but also contributes to reduced power for testing. The non-centrality parameter for testing $H_0 : \beta_x = 0$ with surrogate measurements is $n\beta_x^2\sigma_x^2\rho_{xw}^2/\left\{\sigma_\epsilon^2+\beta_x^2\sigma_x^2\left(1-\rho_{xw}^2\right)\right\}$ which is less than the true-date non-centrality parameter, $n\beta_x^2\sigma_x^2/\sigma_\epsilon^2$, whenever $\rho_{xw}^2 < 1$. These expressions give rise to the equivalent-power sample size formula

$$n_w = n_x\left[\left\{\sigma_\epsilon^2+\beta_x^2\sigma_x^2\left(1-\rho_{xw}^2\right)\right\}/\left\{\sigma_\epsilon^2\rho_{xw}^2\right\}\right]\approx n_x/\rho_{xw}^2, \tag{33.3}$$

where n_w is the number of (W, Y) pairs required to give the same power as a sample of size n_x of (X, Y) pairs. The latter approximation is reasonable near the null value $\beta_x = 0$ (or more precisely, when $\beta_x^2\sigma_x^2\left(1-\rho_{xw}^2\right)$ is small).

The loss of power for testing is not always due to an increase in variability of the parameter estimates. For the classical error model, the variance of the naive estimator is *less than* the variance of the true-data estimator asymptotically if and only if $\beta_x^2\sigma_x^2/(\sigma_x^2+\sigma_u^2) < \sigma_\epsilon^2/\sigma_x^2$, which is possible when σ_ϵ^2 is large, or σ_u^2 is large, or $|\beta_x|$ is small. So relative to the case of no measurement error, classical errors can result in more precise estimates of the wrong (i.e., biased) quantity. This cannot occur with Berkson errors, for which asymptotically the variance of the naive estimator is never less than the variance of the true-data estimator.

The normal linear model also illustrates the need for additional information in measurement error models. For example, for the case of an arbitrary surrogate, the joint distribution of Y and W contains eight unknown parameters $(\beta_1,\ \beta_x,\ \mu_x,\ \mu_w,\ \sigma_x^2,\ \sigma_\epsilon^2,\ \sigma_{xw},\ \sigma_w^2)$, whereas a bivariate normal distribution is completely determined by only five parameters. This means that not all eight parameters can be estimated with data on (Y, W) alone. In particular, β_x is not estimable. However, from Table 33.1, it is apparent that if a consistent estimator of σ_{xw} can be constructed, say from validation data, then the method-of-moments estimator $\widehat{\beta}_x = \left(s_w^2/\widehat{\sigma}_{xw}\right)\widehat{\beta}_w$ is a consistent estimator of β_x, where $\widehat{\beta}_w$ is the least squares estimator of slope in the linear regression of Y on W, s_w^2 is the sample variance of W, and $\widehat{\sigma}_{xw}$ is the validation data estimator of σ_{xw}.

For the case of additive, unbiased, measurement error, the joint distribution of Y and W contains six unknown parameters $(\beta_1,\ \beta_x,\ \mu_x,\ \sigma_x^2,\ \sigma_\epsilon^2,\ \sigma_u^2)$, so that again not all of the parameters are identified. Once again, β_x is not estimable. However, if a consistent estimator of σ_u^2 can be constructed, say from either replicate measurements or validation data, then the method-of-moments estimator $\widehat{\beta}_x = \left\{s_w^2/\left(s_w^2 - \widehat{\sigma}_u^2\right)\right\}\widehat{\beta}_w$ is a consistent estimator of β_x, where $\widehat{\sigma}_u^2$ is the estimator of σ_u^2.

For the Berkson error model, there are also six unknown parameters in the joint distribution of Y and W, $(\beta_1,\ \beta_x,\ \mu_x,\ \sigma_x^2,\ \sigma_\epsilon^2,\ \sigma_w^2)$, so that again not all of the parameters are identified. The regression parameters β_1 and β_x are estimated unbiasedly by the intercept and slope estimators from the least squares regression of Y and W. However, without additional data, it is not possible to estimate σ_ϵ^2.

33.2.4 Multiple Linear Regression

The entries in Table 33.1 and the qualitative conclusions based on them generalize to the case of multiple linear regression with multiple predictors measured with error. For the Berkson error model, it remains the case that no bias in the regression parameter estimators results from the substitution of W for X, and the major effects of measurement error are those resulting from an increase in the residual variation.

For the classical measurement error model, there are important aspects of the problem that are not present in the simple linear regression model. When the model includes both covariates measured with error X and without error Z, it is possible for measurement error to bias the naive estimator of β_z as well as the naive estimator of β_x. Furthermore, attenuation in the coefficient of a variable measured with error is no longer a simple function of the variance of that variable and the measurement error variance. When there are multiple predictors measured with error, the bias in regression coefficients is a non-intuitive function of the measurement error covariance matrix and the true-predictor covariance matrix.

Suppose that the multiple linear regression model for Y given Z and X is $Y = \beta_1 + \beta_z^T Z + \beta_x^T X + \epsilon$. For the additive error model $W = X + U$, the naive estimator of the regression coefficients is estimating

$$\begin{pmatrix}\beta_{z*}\\ \beta_{x*}\end{pmatrix} = \begin{pmatrix}\sigma_{zz} & \sigma_{zx}\\ \sigma_{xz} & \sigma_{xx} + \sigma_{uu}\end{pmatrix}^{-1} \begin{pmatrix}\sigma_{zz} & \sigma_{zx}\\ \sigma_{xz} & \sigma_{xx}\end{pmatrix} \begin{pmatrix}\beta_z\\ \beta_x\end{pmatrix} \tag{33.4}$$

and not $(\beta_z^T, \beta_x^T)^T$. For the case of multiple predictors measured with error with no restrictions on the covariance matrices of the predictors or the measurement errors, bias in individual coefficients can take almost any form. Coefficients can be attenuated toward the null or inflated away from zero. The bias is not always multiplicative. The sign of coefficients can change, and zero coefficients can become non-zero (i.e., null predictors can appear to be significant). There is very little that can be said in general, and individual cases must be analyzed separately.

However, in the case of one variable measured with error, that is, X is a scalar, the attenuation factor in β_{x*} is $\lambda_1 = \sigma^2_{x|z}/(\sigma^2_{x|z} + \sigma^2_u)$ where $\sigma^2_{x|z}$ is the residual variance from the regression of X on Z, that is, $\beta_{x*} = \lambda_1\beta_x$. Because $\sigma^2_{x|z} \leq \sigma^2_x$, attenuation is accentuated relative to the case of no covariates when the covariates in the model are correlated with X, that is, $\lambda_1 \leq \lambda$ with strict inequality when $\sigma^2_{x|z} < \sigma^2_x$. Also, in the case of a single variable measured with error, $\beta_{z*} = \beta_z + (1 - \lambda_1)\beta_x\Gamma_z$, where Γ_z is the coefficient vector of Z in the regression of X on Z, that is, $E(X \mid Z) = \Gamma_1 + \Gamma_z^T Z$. Thus, measurement error in X can induce bias in the regression coefficients of Z. This has important implications for analysis of covariance models in which the continuous predictor is measured with error (Carroll 1989; Carroll et al. 1985).

The effects of measurement error on naive tests of hypotheses can be understood by exploiting the fact that in the classical error model, W is a surrogate. In this case, $E(Y|Z,W) = E\{E(Y|Z,X,W)|Z,W\} = E\{E(Y|Z,X)|Z,W\} = \beta_1 + \beta_z^T Z + \beta_x^T E(X|Z,W)$. With multivariate normality, $E(X|Z,W)$ is linear, say $E(X|Z,W) = \alpha_0 + \alpha_z^T Z + \alpha_w W$, and thus

$$E(Y|Z,W) = \beta_0 + \beta_x^T\alpha_0 + \left(\beta_z^T + \beta_x^T\alpha_z^T\right) Z + \beta_x^T\alpha_w^T W. \tag{33.5}$$

This expression holds for any surrogate W. Our summary of hypothesis testing in the presence of measurement error is appropriate for any surrogate variable model provided α_w^T is an invertible matrix, as it is for the classical error model. Suppose that the naive model is parameterized

$$E(Y|Z,W) = \gamma_0 + \gamma_z^T Z + \gamma_w^T W. \tag{33.6}$$

A comparison of (33.5) and (33.6) reveals the main effects of measurement error on hypothesis testing.

First note that $(\beta_z^T, \beta_x^T)^T = 0$ if and only if $(\gamma_z^T, \gamma_x^T)^T = 0$. This implies that the naive-model test that none of the predictors are useful for explaining variation in Y is valid in the sense of having the desired type I error rate. Further examination of (33.5) and (33.6) shows that $\gamma_z = 0$ is equivalent to $\beta_z = 0$, only if $\alpha_z\beta_x = 0$. It follows that the naive test of $H_0 : \beta_z = 0$ is valid only if X is unrelated to Y ($\beta_x = 0$) or if Z is unrelated to X ($\alpha_z = 0$). Finally, the fact that $\beta_x = 0$ is equivalent to $\alpha_w\beta_x = 0$ implies that the naive test of $H_0 : \beta_x = 0$ is valid. The naive tests that are valid, that is, those that maintain the type I error rate, will still suffer reduced power relative to the test based on the true data.

33.2.5 Non-Linear Regression

The effects of measurement error in non-linear models are much the same qualitatively as in the normal linear model. The use of a surrogate measurement generally results in reduced power for testing associations, produces parameter bias, and results in a model with less predictive power. However, the nature of the bias depends on the model, the type of parameter, and the error model. Generally, the more non-linear the model is, the less relevant are the results for the linear model. Parameters other than linear regression coefficients (e.g., polynomial coefficients, transformation parameters, and variance function parameters) have no counterpart in the normal linear model and the effect of measurement errors on such parameters must be studied on a case-by-case basis.

Regression coefficients in generalized linear models, including models of particular interest in epidemiology such as logistic regression and Poisson regression, are affected by measurement error in much the same manner as are linear model regression coefficients. This means that relative risks and odds ratios derived from logistic regressions models are affected by measurement error much the same as linear model regression coefficients (Rosner et al. 1989, 1990; Stefanski 1985; Stefanski and Carroll 1985). However, unlike the linear model, unbiased Berkson measurements generally produce biases in non-linear models, although they are often much less severe than biases resulting from classical measurement errors (for comparable ρ_{xw}). This fact forms the basis for the method known as *regression calibration* in which an unbiased Berkson predictor is estimated by a preliminary calibration analysis, and then the usual (naive) analysis is performed with $\widehat{E(X|W)}$ replacing X. This fact also explains why more attention is paid to the classical error model than to the Berkson error model.

The effects of classical measurement error on flexible regression models, for example, non-parametric regression, are not easily quantified, but there are general tendencies worth noting. Measurement error generally smooths out regression functions. Non-linear features of $E(Y|X)$ such as curvature of local extremes, points of non-differentiability, and discontinuities will generally be less pronounced or absent in $E(Y|W)$. For normal measurement error, $E(Y|W)$ is smooth whether $E(Y|X)$ is or is not, and local maxima and minima will be less extreme – measurement error tends to wear off the peaks and fill in the valleys. This can be seen in a simple parametric model. If $E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2$ and (X, W) are jointly normal with $\mu_x = 0$, then $E(Y|W)$ is also quadratic with the quadratic coefficient attenuated by ρ_{xw}^4. The local extremes of the two regressions differ by $\beta_2 \sigma_x^2 (1 - \rho_{xw}^2)$ which is positive (negative) when $E(Y|X)$ is convex (concave). Finally, we note that monotonicity of regression functions can sometimes be affected by heavy-tailed measurement error (Hwang and Stefanski 1994).

The effects of classical measurement error on density estimation are qualitatively similar to that of non-parametric regressions. Modes are attenuated and regions of low density are inflated. Measurement error can mask multimodality in the true density and will inflate the tails of the distribution. Naive estimates of tail quantiles are generally more extreme than the corresponding true-data estimates.

33.2.6 Logistic Regression Example

This section closes with an empirical example illustrating the effects of measurement error in logistic regression, and the utility of the multivariate normal linear regression model results for approximating the effects of measurement error. The data used are a subset of the Framingham Heart Study data and are described in detail in Carroll et al. (2006). For these data, X is long-term average systolic blood pressure after transformation via ln(SBP-50), denoted TSBP. There are replicate measurements (W_1, W_2) for each of $n = 1,615$ subjects in the study. The true-data model is logistic regression of coronary heart disease (CHD) $(0, 1)$ on X and covariates (Z) including age, smoking status (SMOKE) $(0, 1)$, and cholesterol level (CHOL).

Assuming the classical error model for the replicate measurements, $W_j = X + U_j$, analysis of variance produces the estimate $\widehat{\sigma}_u^2 = 0.0126$. The average $\overline{W} = (W_1 + W_2)/2$ provides the best measurement of X with an error variance of $\sigma_U^2/2$ (with estimate 0.0063).

The three measurements, W_1, W_2 and $\overline{W}$, can be used to empirically demonstrate attenuation due to measurement error. The measurement error variances of W_1 and W_2 are equal and are twice as large the measurement error variance of $\overline{W}$. Thus, the attenuation in the regressions using W_1 and W_2 should be equal, whereas the regression using $\overline{W}$ should be less attenuated. Three naive logistic models,

$$\text{logit}\{\Pr(\text{CHD=1})\} = \beta_0 + \beta_{z_1}\text{AGE} + \beta_{z_2}\text{SMOKE} + \beta_{z_3}\text{CHOL} + \beta_x\text{TSBP},$$

were fit using each of the three measurements W_1, W_2, and $\overline{W}$. The estimates of the TSBP coefficient from the logistic regressions using W_1 and W_2 are both 1.5 (to one decimal place). The coefficient estimate from the fit using $\overline{W}$ is 1.7. The relative magnitudes of the coefficients $(1.5 < 1.7)$ are consistent with the anticipated effects of measurement error – greater attenuation associated with larger error variance. The multiple linear regression attenuation coefficient for a measurement with error variance σ^2 is $\lambda_1 = \sigma_{x|z}^2/(\sigma_{x|z}^2 + \sigma^2)$. Assuming that this applies approximately to the logistic model suggests that

$$1.7 \approx \frac{\sigma_{x|z}^2}{\sigma_{x|z}^2 + \sigma_u^2/2}\beta_x \qquad \text{and} \qquad 1.5 \approx \frac{\sigma_{x|z}^2}{\sigma_{x|z}^2 + \sigma_u^2}\beta_x.$$

Because β_x is unknown, these approximations cannot be checked directly. However, a check on their consistency is obtained by taking ratios leading to $1.13 = 1.7/1.5 \approx (\sigma_{x|z}^2 + \sigma_u^2)/(\sigma_{x|z}^2 + \sigma_u^2/2)$. Using the ANOVA estimate, $\widehat{\sigma}_u^2 = 0.0126$, and the mean squared error from the linear regression of $\overline{W}$ on AGE, SMOKE, and CHOL as an estimate of $\sigma_{\overline{w}|z}^2$, produces the estimate $\widehat{\sigma}_{x|z}^2 = \widehat{\sigma}_{\overline{w}|z}^2 - \widehat{\sigma}_u^2/2 = 0.0423 - 0.0063 = 0.0360$. Thus, $(\sigma_{x|z}^2 + \sigma_u^2)/(\sigma_{x|z}^2 + \sigma_u^2/2)$ is estimated to be $(0.0360 + 0.0126)/(0.0360 + 0.0063) = 1.15$. In other words, the attenuation in the logistic regression coefficients is consistent ($1.13 \approx 1.15$) with the attenuation predicted by the normal linear regression model result.

These basic statistics can also be used to calculate a simple bias-adjusted estimator as $\widehat{\beta}_x = 1.7(\widehat{\sigma}^2_{x|z} + \widehat{\sigma}^2_u/2)/\widehat{\sigma}^2_{x|z} = 1.7(0.0360 + 0.0063)/0.0360 = 2.0$, which is consistent with estimates reported by Carroll et al. (2006) obtained using a variety of measurement error estimation techniques. We do not recommend using linear model corrections for the logistic model because there are number of methods more suited to the task as described in Sect. 33.4. Our intent with this example is to demonstrate the general relevance of the easily derived theoretical results for linear regression to other generalized linear models.

The odds ratio for a Δ change in transformed systolic blood pressure is $\exp(\beta_x \Delta)$. With the naive analysis, this is estimated to be $\exp(1.7\Delta)$; the bias-corrected analysis produces the estimate $\exp(2.0\Delta)$. Therefore, the naive odds ratio is attenuated by approximately $\exp(-0.3\Delta)$. More generally, the naive (OR_N) and true (OR_T) odds ratios are related via $OR_N/OR_T = OR_T^{\lambda_1 - 1}$, where λ_1 is the attenuation factor in the naive estimate of β_x. The naive and true relative risks have approximately the same relationship under the same conditions (small risks) that justify approximating relative risks by odds ratios.

33.3 Planning Epidemiological Studies with Measurement Error

As the previous sections have established, exposure measurement error is common in epidemiological studies and, under certain assumptions, can be shown to have dramatic effects on the properties of relative risk estimates or other types of coefficients derived from epidemiological regression models. It is wise therefore to include measurement error considerations in the planning of a study, both to enable the application of a measurement error analysis at the conclusion of the study and to assure scientific validity.

In developing a useful plan, one must consider a number of important questions. To begin with, what are the scientific objectives of the study? Is the goal to identify a new risk factor for disease, perhaps for the first time, or is this a study to provide improved estimates of the quantitative impact of a known risk factor? Is prediction of future risks the ultimate goal? Do the measurement errors affect adjusting covariates as well? The answers to these questions will determine the possible responses to dealing with the measurement error in the design and analysis of the study, including the choice of a criterion for statistical optimality. It is even possible that no measurement error correction is needed to achieve the purposes of the study (Thiebaut et al. 2007), or in certain instances, absent other considerations such as cost, that *the most scientifically valid design would eliminate measurement error entirely.*

The nature of the measurement error should be carefully considered. For instance, is the measurement error non-differential? What is the evidence to support this conclusion? Especially in the study of complex phenomenon such as nutritional factors in disease, the non-differential assumption deserves scrutiny. For example, the diet "record" has often been used as the gold standard of nutritional intakes, but subsequent analyses have cast doubt on the non-differential measurement error associated with substituting monthly food frequency questionnaires (Kipnis et al. 1999). On the other hand, measurement errors due to validated scientific instrument errors may be more easily justified as non-differential.

Another thing to consider is the possible time dependency of exposure errors, and how this may affect the use of non-differential models. This often arises in case-control studies where exposures must be assessed retrospectively (cf. chapter ▸Case-Control Studies of this handbook). An interesting example occurs in a recent study of arsenic exposure where both drinking water and toenail measurements are available as personal exposure measures in a cancer case-control study (Karagas et al. 1998). Toenail concentrations give a biologically time-averaged measure of exposure, but the time scale is limited and the nail concentrations are influenced by individual metabolic processes. Drinking water concentrations may be free from possible confounding due to unrelated factors affecting metabolic pathways but could be less representative of average exposures over the time interval of interest. This kind of ambiguity is common in many epidemiological modeling situations and should indicate caution in the rote application of measurement error methods.

Depending on the type of non-differential error, different study plans may be required to identify the desired relative risk parameters. For instance, replicate measurements of an exposure variable may adequately identify the necessary variance parameters in a classical measurement error model. Under certain circumstances, an "instrumental" variable may provide the information needed to correct for measurement error. This type of reliability/validity data leads to identifiable relative risk regression parameters in classical or Berkson case error.

In more complex "surrogate" variable situations with non-differential error, an internal or external validation study may be necessary, where the "true" exposure measured without error is available for a subset or independent sample of subjects. These designs are also useful and appropriate for classical measurement error models but are essential in the case of surrogates which cannot be considered "unbiased." Internal validation studies have the capability of checking the non-differential assumption and thus are potentially more valuable. With external validation studies, there may be doubt as to whether the populations characterized by the validation and main study samples are comparable in the sense that the measurement error model is equivalent or "transportable" between the populations. The issue of whether an error model is transportable or not can arise with any type of measurement error (Table 33.2).

The considerations described above are summarized in the following table for some of the options that should be considered when planning a study in the presence of measurement error.

Because internal validation studies involve the observation of the true exposure and the health outcome for a portion of the sample, this design has a distinct

Table 33.2 Sampling plan characteristics for collecting validation data in epidemiological studies with measurement error of different types or properties

	Validation data			
Measurement error type/property	Replicates	Instrumental variables	External study	Internal study
Classical	Yes	Yes	Yes	Yes
Berkson	No	Yes	Yes	Yes
General surrogate	No	No	Yes	Yes
Differential	No	No	No	Yes

scientific appeal. However, relative costs will drive the choice of an optimal design. For instance, it may be the case that a classical additive error model applies and that replicate measures are easier or cheaper to get than true values. Depending on the relative impact on the optimality criterion used, the replicate design might be more cost-effective.

A number of approaches have been suggested and used for the design of epidemiological studies based on variables measured with error. These may be characterized broadly as sample size calculation methods, where the design decision to be made has to do mainly with the size of the main study in studies where the measurement error is known or can be ignored, and design approaches for studies using internal or external validation data where both the size of the main study and the validation sample must be chosen. In the sections that follow, we review both of these approaches.

33.3.1 Methods for Sample Size Calculations

Methods for sample size calculations are typically based on the operating characteristics of a simple hypothesis test. In the case of measurement error in a risk factor included in an epidemiological regression model, the null hypothesis is that the regression coefficient for the risk factor equals zero, implying no association between the exposure and the health outcome. For a specific alternative, one might calculate the power for a given sample size or, alternatively, the sample size required to achieve a given power.

It has been known for some time that the effect of measurement error is to reduce the power of the test for no association both in linear models (Cochran 1968) and 2×2 tables with non-differential misclassification (Fleiss 1981). This result has been extended to survival models (Prentice 1982) and to generalized linear models with non-differential exposure measurement error (Tosteson and Tsiatis 1988), including linear regression, logistic regression, and tests for association in 2×2 contingency tables. Using small relative risk approximations, it is possible to show that for all of these common models for epidemiological data, the ratio of the sample size required using the data measured without error to the sample size required using the error prone exposure is approximately $n_x/n_w \approx \rho^2_{xw}$, the square of the correlation between X and W; see also Eq. 33.3, Sect. 33.2.3. This relation provides a handy method for determining sample size requirements in the presence of measurement error as

$$n_w = n_x/\rho^2_{xw}. \tag{33.7}$$

If additional covariates Z are included in the calculation, a partial correlation can be used instead. The same formula has been used for sample size calculations based on regression models for prospective studies with log-linear risk functions and normal distributions for exposures and measurement error (McKeown-Eyssen and Tibshirani 1994) and case-control studies with conditionally normal exposures within the case and control groups (White et al. 1994). Recent developments have improved this approximation (Tosteson et al. 2003), but Formula (33.7) remains a useful tool for checking sample size requirements in studies with measurement error.

For generalized linear models (Tosteson and Tsiatis 1988) and survival models (Prentice 1982), it has been shown that optimal score test can be computed by replacing the error prone exposure variable W with $E[X|W]$, a technique that was later termed regression calibration (Carroll et al. 2006). Subsequent work extended these results to a more general form of the score test incorporating a non-parametric estimate of the measurement error distribution (Stefanski and Carroll 1990a). One implication of this result is that in common measurement error models, including normally distributed exposure errors and non-differential misclassification errors, the optimal test is computed simply by ignoring the measurement error and using the usual test based on W rather than X, the true exposure. However, the test will still suffer the loss of power implicit in Formula (33.7).

It is interesting to consider the effects of Berkson case errors on sample size calculations. The implication for analysis is somewhat different, in as much as regression coefficients are unbiased by Berkson case errors for linear models and to the first order for all generalized linear models. However, as applied to epidemiological research, there is no distinction with respect to the effects of this type of non-differential sample size calculations for simple regression models without confounders, and Formula (33.7) applies directly.

33.3.2 Planning for Reliability/Validation Data

In some epidemiological applications, collection of data for a measurement error correction will be planned, although this may be deemed unnecessary in some simple situations where the investigators only wish to demonstrate an association with an error-prone exposure and with no error in covariates or where the measurement errors are known. Information on the measurement error parameters can come from a number of possible designs, including replicate measurements, instrumental variables, external validation studies measuring the true and surrogate exposures (i.e., just X and W), or internal validation studies. A variety of statistical criteria can be used to optimize aspects of the design, most commonly the variance of the unbiased estimate of the relative risk for the exposure measured with error. Other criteria have included the power of tests of association, as in the previous section, and criteria based on the power of tests for null hypotheses other than no association (Spiegelman and Gray 1991).

To choose a design, it is usually necessary to have an estimate of the measurement error variance or other parameters. This may be difficult since validation data are needed to derive these estimates and will not yet have been collected at the time when the study is being planned. However, this dilemma is present in most practical design settings and can be overcome in a number of informal ways by deriving estimates from previous publications, pilot data, or theoretical considerations of the measurement error process. Certain sequential designs can be useful in this regard, and some suggestions are discussed here in the context of the design of internal validation studies.

In studies where a correction is planned for classical measurement error using replicates, the simple approach to sample size calculations may provide a guideline for choosing an appropriate number of replicates and a sample size by replacing ρ^2_{xw} with $\rho^2_{x\overline{w}}$, where $\overline{w}$ is the mean of the n_r replicates. Depending on the relative costs of replication and obtaining a study participant, these expressions may be used to find an optimal value for the overall sample size, n, and the number of replicates, n_r. For instrumental variables, a similar calculation can be made using a variation on the regression calibration procedure as applied to the score test for no association. In this case, the inflation in sample size for (33.7) is based on $\rho^2_{\widehat{x}x}$, where $\widehat{x} = E[X|W, T]$, the predicted value of the true exposure given the unbiased surrogate W and the instrumental variable T.

External and internal validation studies both involve a main study, with a sample size of n_1 and a validation study, with sample size of n_2. The external validation study involves an independent set of measurements of the true and surrogate exposures, whereas the internal validation study is based on a subset of the subjects in the main study. Both the size of the main study and the validation study must be specified. In the internal validation study, n_2 is by necessity less than or equal to n_1, with equality implying a fully validated design. In the external validation study, n_2 is not limited, but the impact of increasing the amount of validation data is more limited than in the internal validation study. This is because the fully validated internal validation study has no loss of power versus a study that has no measurement error, whereas the external validation study can only improve the power to the same as that of a study with measurement error where the measurement error parameters are known.

For common non-linear epidemiological regression analyses such as logistic regression, method for determining optimal values of n_1 and n_2 has typically involved specialized calculations (Spiegelman and Gray 1991; Stram et al. 1995). Less intractable expressions are available for linear discriminant models, not involving numerical integrations (Buonaccorsi 1988). The actual analysis of the data from the studies may be possible using approximations such as the regression calibration method requiring less sophisticated software (Spiegelman et al. 2001).

A variant on the internal validation study is designs which use surrogate exposures and outcomes as stratification variables to select a highly efficient validation sample. Cain and Breslow (1988) develop methods for case-control studies where surrogate variables were available during the design phase for cases and controls. Tosteson and Ware (1990) develop methods for studies where surrogates were available for both exposures and a binary outcome. These designs can be analyzed with ordinary logistic regression if that model is appropriate for the population data. Methods for improving the analysis of the designs and adapting them to other regression models have been proposed (Tosteson et al. 1994; Holcroft et al. 1997; Reilly 1996).

33.3.3 Examples and Applications

Much of the research on methods for planning studies with measurement error has been stimulated by applications from environmental, nutritional, and occupational epidemiology. Nevertheless, it is fair to say that published examples of studies designed with measurement error in mind are relatively rare, and the best source of case studies may be method papers such as those cited in this review. This may reflect a lack of convenient statistical software other than what individual researchers have been able to make available.

One example of a study using planning based on measurement error methods was provided by IJmker et al. (2006). The exposure of interest was the intensity of use of computers as measured by worker monitoring software, and the outcomes are musculoskeletal disorders. A web-based program for logistic regression based on the methods of Tosteson et al. (2003) was used to adjust the sample size for possible exposure measurement error.

Even without widely distributed software, some useful calculations can be quite simple, as shown above, and a more important factor in future applications of these methods will be proper education to raise the awareness among statisticians and epidemiologists of the importance of addressing the problem of measurement error in the planning phases of health research.

33.4 Measurement Error Models and Methods

33.4.1 Overview

This section describes several common methods for correcting biases induced by non-differential covariate measurement error. The focus is on non-linear regression models and the logistic model in particular, though all the methods apply to the linear model. The intent is to familiarize the reader with the central themes and key ideas that underlie the proposals and provide a contrast of the assumptions and types of data required to implement the procedures. The literature contains a very large number of proposals for measurement error correction in regression models, and the presentation here is necessarily focused. Texts that provide broad and detailed discussions of modern methods for handling measurement error include Carroll et al. (2006), Buonaccorsi (2010), and Gustafson (2004).

The starting point for all measurement error analyses is the disease model of interest relating the disease outcome Y to the true exposure(s) X and covariates Z and a measurement error model relating the mismeasured exposure W to (Z, X). Measurement error methods can be grouped according to whether they employ *functional* or *structural* modeling. Functional models make no assumptions on X, beyond what are made in the absence of measurement error, for example, $\sum_{i=1}^{n}(X_i - \overline{X})^2 > 0$ for simple linear regression (the use of the word *functional* here does not refer to the area of functional data analysis). Functional modeling is compelling because often there is little information in the data on the distribution of X. For this reason, much of the initial research in measurement error methods focused on functional modeling. Methods based on functional modeling can be divided into approximately consistent (remove most bias) and fully consistent methods (remove all bias as $n \rightarrow \infty$). Fully consistent methods for non-linear regression models often require assumptions on the distribution of the measurement error. Regression calibration and SIMEX are examples of approximately consistent methods while corrected scores, conditional scores, Tsiatis and Ma scores (Tsiatis and Ma 2004), and some instrumental variable (IV) methods are fully consistent for large classes of models. Each of these approaches is described below.

Structural models assume that the unobserved exposure X is random and require an exposure model for X, with the normal distribution as the default model. Bayesian and likelihood-based methods are used with structural models.

Note that the terms functional and structural refer to assumptions on X, not on the measurement error model. Functional modeling has the advantage of providing valid inference regardless of the distribution of X. On the other hand, structural modeling can result in large gains in efficiency and allows construction of likelihood ratio based confidence intervals that often have coverage probabilities closer to the nominal level than large sample normal theory intervals used with functional models. The choice between functional or structural modeling depends both on the assumptions one is willing to make and, in a few cases, the form of the model relating Y to (Z, X). The type and amount of data available also play a role. For example, validation data provides information on the distribution of X and may make structural modeling more palatable. The remainder of this chapter describes methods for correcting for measurement error. Functional methods are described first.

33.4.2 Regression Calibration

Regression calibration is a conceptually straightforward approach to bias reduction and has been successfully applied to a broad range of regression models. It is the default approach for the linear model. The method is fully consistent in linear models and log-linear models when the conditional variance of X given (Z, W) is constant. Regression calibration is approximately consistent in non-linear models. The method was first studied in the context of proportional hazards regression (Prentice 1982). Extensions to logistic regression and a general class of regression models were studied in Rosner et al. (1989, 1990) and Carroll and Stefanski (1990), respectively. A detailed and comprehensive discussion of regression calibration can be found in Carroll et al. (2006).

When the measurement error is non-differential, the induced disease model, or regression model, relating Y to the observed exposure W and covariates Z is $E[Y \mid Z, W] = E[E[Y \mid Z, X] \mid Z, W]$, that is, the induced disease model is obtained by regressing the true disease model on (Z, W). A consequence of the identity is that the form of the observed disease model depends on the conditional distribution of X given (Z, W). This distribution is typically not known, and even when known, evaluating the right-hand side of the identity can be difficult. For example, if the true disease model is logistic and the distribution of X conditional on (Z, W) is normal, there is no closed form expression for $E[Y \mid Z, W]$.

Regression calibration circumvents these problems by approximating the disease model relating Y to the observed covariates (Z, W). The approximation is obtained by replacing X with $E[X \mid Z, W]$ in the model relating Y to (Z, X). Because regression calibration provides a model for Y on (Z, W), the observed data can be used to assess the adequacy of the model.

To describe how to implement the method, it is useful to think of the approach as a method for imputing values for X. The idea is to estimate unobserved X with $X^* \equiv$ predicted value of X from the regression of X on (Z, W); see the discussion of Berkson error calibration in Sect. 33.2.2. Modeling and estimating the regression of X on (Z, W) requires additional data in the form of internal/external replicate observations, instrumental variables, or validation data; see the example below. The regression parameters in the true disease model are estimated by regressing Y on (Z, X^*). Note that X^* is the best estimate of X using the observed predictors (Z, W), best in the sense of minimizing mean square prediction error. To summarize, regression calibration estimation consists of two steps:

1. Model and estimate the regression of X on (Z, W) to obtain X^*.
2. Regress Y on (Z, X^*) to obtain regression parameter estimates.

A convenient feature of regression calibration is that standard software can often be used for estimation. However, standard errors for parameter estimates in Step 2 must account for the fact that X^* is estimated in Step 1, something standard software does not do. Bootstrap or asymptotic methods based on estimating equation theory are typically used; see Carroll et al. (2006) for details.

When (Z, X, W) is approximately jointly normal, or when X is strongly correlated with (Z, W), the regression of X on (Z, W) is approximately linear:

$$E[X \mid Z, W] \approx \mu_x + \sigma_{x|zw}\sigma_{zw}^{-1}\begin{pmatrix} Z - \mu_z \\ W - \mu_w \end{pmatrix}$$

where $\sigma_{x|zw}$ is the covariance of X with (Z, W) and σ_{zw} is the variance matrix of (Z, W). Implementing regression calibration using the linear approximation requires estimation of the calibration parameters μ_x, $\sigma_{x|zw}$, σ_{zw}, μ_w, and μ_z.

33.4.2.1 Example

We illustrate estimation of the calibration function when two replicate observations of X are available in the primary study (internal reliability data) and the error model is $W = X + \sigma_u U$. For ease of illustration, we assume there are no additional covariates Z. Let $\{W_{i1}, W_{i2}\}_{i=1}^{n}$ denote the replication data and suppose that $E[X_i \mid \overline{W}_i] \approx \mu_x + \sigma_{x|\overline{w}}\sigma_{\overline{w}}^{-1}(\overline{W}_i - \mu_w) = \mu_w + \frac{\sigma_{\overline{w}}^2 - \sigma_u^2/2}{\sigma_{\overline{w}}^2}(\overline{W}_i - \mu_w)$ where $\overline{W}_i = (W_{i1} + W_{i2})/2$, and the last equality follows from the form of the error model. Note that $\frac{\sigma_{\overline{w}}^2 - \sigma_u^2/2}{\sigma_{\overline{w}}^2}$ is the attenuation factor introduced in Sect. 33.2.3. The method-of-moments calibration parameter estimators are $\widehat{\mu}_w = \sum_{i=1}^{n} \overline{W}_i/n$, $\widehat{\sigma}_{\overline{w}}^2 = \sum_{i=1}^{n}(\overline{W}_i - \widehat{\mu}_w)^2/(n-1)$ and $\widehat{\sigma}^2 = \sum_{i=1}^{n}\sum_{j=1}^{2}(W_{ij} - \overline{W}_i)^2/n = \sum_{i=1}^{n}(W_{i1} - W_{i2})^2/2n$. The imputed value for X_i is $X_i^* = \widehat{\mu}_w + \frac{\widehat{\sigma_{\overline{w}}^2} - \widehat{\sigma_u^2}/2}{\widehat{\sigma_{\overline{w}}^2}}(\overline{W}_i - \widehat{\mu}_w)$.

For the Framingham data described in Sect. 33.2.6, recall that $\{W_{i1}, W_{i2}\}_{i=1}^{1{,}615}$ represented transformed systolic blood pressure measured for each subject at two separate exams. For these data, $\widehat{\sigma}_u^2 = 0.0126, \widehat{\sigma}_{\overline{w}}^2 = 0.0454$ and $\widehat{\mu}_w = 4.36$ so that the imputed measurement is $X_i^* = 4.36 + .86(\overline{W}_i - 4.36)$.

If the model relating Y to X is the simple linear regression model, ($Y = \beta_1 + \beta_x X + \epsilon$), regressing Y on X^* results in $\widehat{\beta}_x = \frac{\widehat{\sigma_{\overline{w}}^2}}{\widehat{\sigma_{\overline{w}}^2} - \widehat{\sigma_u^2}/2}\widehat{\beta}_{\overline{w}}$ where $\widehat{\beta}_{\overline{w}}$ is the naive estimator obtained from regressing Y on $\overline{W}$. Note for the linear model the regression calibration estimator coincides with the method-of-moments estimator given in Sect. 33.2 of this chapter.

Our illustration of calibration parameter estimation assumed exactly that two replicates were available for each X_i. This estimation scheme can be easily extended to an arbitrary number of replicates for each X_i (Carroll et al. 2006).

33.4.2.2 Additional Notes

Regression calibration can be ineffective in reducing bias in non-linear models when (a) the effect of X on Y is large, for example, large odds ratios in logistic regression, (b) the measurement error variance is large, and (c) the model relating Y to (Z, X) is not smooth. It is difficult to quantify what is meant by large in (a) and (b) because all three factors (a–c) can act together. Segmented regression is an example of a model where regression calibration fails due to lack of model smoothness (Küchenhoff and Carroll 1997). Segmented models relate Y to X using separate regression models on different segments along the range of X.

In logistic regression, the method has been found to be effective in a number of applications (Rosner et al. 1989, 1990; Carroll et al. 2006). Fearn et al. (2008) provide a recent epidemiological application of regression calibration. Extensions for regression calibration that are designed to address the potential pitfalls listed in (a)–(c) are given in Carroll and Stefanski (1990).

Regression calibration has been implemented for generalized linear models in STATA.

33.4.3 SIMEX

Simulation extrapolation (SIMEX) is a very broadly applicable method for bias correction and is the only method that provides a visual display of the effects of measurement error on regression parameter estimation. SIMEX is fully consistent for linear disease models and approximate for non-linear models. SIMEX is founded on the observation that bias in parameter estimation varies in a systematic way with the magnitude of the measurement error. Essentially, the method is to incrementally add measurement error to covariate W using computer-simulated random errors and then compute the corresponding regression parameter estimates (simulation step). The extrapolation step models the relation between the parameter estimates and the magnitude of the measurement errors. The SIMEX estimate is the extrapolation of this relation to the case of zero measurement error.

Details of the method are best understood in the context of the classical additive measurement error model. However, the method is not limited to this model. To describe the method, suppose $W_i = X_i + \sigma_u U_i$ for $i = 1, \ldots, n$ and for $s = 1, \ldots, B$, define $W_{is}(\lambda) = W_i + \sqrt{\lambda}\sigma_u U_{is}$ where $\lambda > 0$, and $\{U_{is}\}_{s=1}^{B}$ are i.i.d. computer-simulated standard normal variates. Note that the variance of the measurement error for the constructed measurement $W_{is}(\lambda)$ is $(1+\lambda)\sigma_u^2$, indicating that λ regulates the magnitude of the measurement error. Let $\widehat{\beta}_s(\lambda_j)$ denote the vector of regression parameter estimators obtained by regression of Y on $\{Z, W_s(\lambda_j)\}$ for $0 = \lambda_1 < \lambda_2 < \cdots < \lambda_M$. The value $\lambda_M = 2$ is recommended (Carroll et al. 2006). The notation explicitly indicates the dependence of the estimator on λ_j. Let $\widehat{\beta}(\lambda_j) = B^{-1}\sum_{s=1}^{B}\widehat{\beta}_s(\lambda_j)$. Here we are averaging over the B simulated samples to eliminate variability due to simulation, and empirical evidence suggests $B = 100$ is sufficient. Each component of the vector $\widehat{\beta}(\lambda)$ is then modeled as a function of λ, and the SIMEX estimator is the extrapolation of each model to $\lambda = -1$. Note that $\lambda = -1$ represents a measurement error variance of zero.

Consider, for example, estimation of β_x. The "observations" produced by the simulation $\{\widehat{\beta}_x(\lambda_j), \lambda_j\}_{j=1}^{M}$ are plotted and used to develop and fit an extrapolation model relating the dependent variable $\widehat{\beta}_x(\lambda)$ to the independent variable λ. In most applications, an adequate extrapolation model is provided by either the non-linear extrapolant function, $\widehat{\beta}_x(\lambda_j) \approx \gamma_1 + \frac{\gamma_2}{\gamma_3+\lambda}$, or a quadratic extrapolant function, $\widehat{\beta}_x(\lambda_j) \approx \gamma_1 + \gamma_2\lambda + \gamma_3\lambda^2$. The appropriate extrapolant function is fit to $\{\widehat{\beta}_x(\lambda_j), \lambda_j\}_{j=1}^{M}$ using ordinary least squares. It is worth noting that the non-linear extrapolant function can be difficult to fit numerically and details for doing so are given in Carroll et al. (2006).

33.4.3.1 Analytical Example

SIMEX was developed to understand and correct for the effects of covariate measurement error in non-linear disease models. However, it is instructive to consider the simple linear regression model to illustrate analytically the relation between $\widehat{\beta}(\lambda)$ and λ. In Sect. 33.2, the bias of the naive estimator was studied and it follows that $\widehat{\beta}_x(\lambda) = \frac{\beta_x \sigma_x^2}{\sigma_x^2 + \sigma_u^2(1+\lambda)} + O_p(n^{-\frac{1}{2}})$ where the symbol $O_p(n^{-\frac{1}{2}})$ denotes terms that are negligible for n large. Therefore, the non-linear extrapolant will result in a fully consistent estimator; $\widehat{\beta}_x(-1) = \frac{\beta_x \sigma_x^2}{\sigma_x^2 + \sigma_u^2(1+[-1])} + O_p(n^{-\frac{1}{2}}) = \beta_x + O_p(n^{-\frac{1}{2}})$.

33.4.3.2 Graphical Example

The Framingham data described in Sect. 33.2.6 are used here to graphically illustrate the SIMEX method. In that section, a logistic model was defined relating the probability of developing coronary heart disease to age, smoking status, cholesterol level, and a transformation of systolic blood pressure. Figure 33.1 depicts the effect of increasing amounts of measurement error on parameter estimates (log odds ratios) and the SIMEX extrapolation to the case of no measurement error. Note that the non-linear and quadratic extrapolants result in similar estimates.

33.4.3.3 Additional Notes

SIMEX has been implemented in the commercial statistical software package STATA and the free open source programming environment R as an R package (simex). Both implementations provide for different choices of the extrapolant function and provide estimates of standard error.

The SIMEX method was developed in a series of papers (Cook and Stefanski 1995; Stefanski and Cook 1995; Carroll et al. 1996) and is summarized in detail in Carroll et al. (2006). Recent epidemiological applications of the method include Kopecky et al. (2006), de Gramont et al. (2007), and Kottgen et al. (2007).

Further refinements and applications of the SIMEX method appear in a number of papers (Stefanski and Bay 1996; Lin and Carroll 1999; Wang et al. 1998; Li and Lin 2003; Kim and Gleser 2000; Kim et al. 2000; Holcomb 1999; Marcus and Elias 1998). Extensions to non-parametric regression are studied in Carroll et al. (1999a), Staudenmayer and Ruppert (2004), and Delaigle and Hall (2008). SIMEX has also been extended to measurement error in categorical covariates; see Küchenhoff et al. (2006).

33.4.4 Estimating Equations and Corrected Scores

Regression parameter estimators in non-linear models are defined implicitly through estimating equations (cf. chapter ▶Generalized Estimating Equations of this handbook). Estimating equations are often based on the likelihood *score*, that is, the derivative of the log-likelihood, or quasi-likelihood scores that only require assumptions on the first and second conditional moments of the disease model. The criterion of least squares also leads to parameter estimation based on estimating equations.

Corrected scores, conditional scores, and certain instrumental variable methods have been developed starting with estimating equations that define regression parameter estimates in the absence of measurement error. An estimating score is *unbiased* if it has expectation zero and is otherwise biased. Measurement error

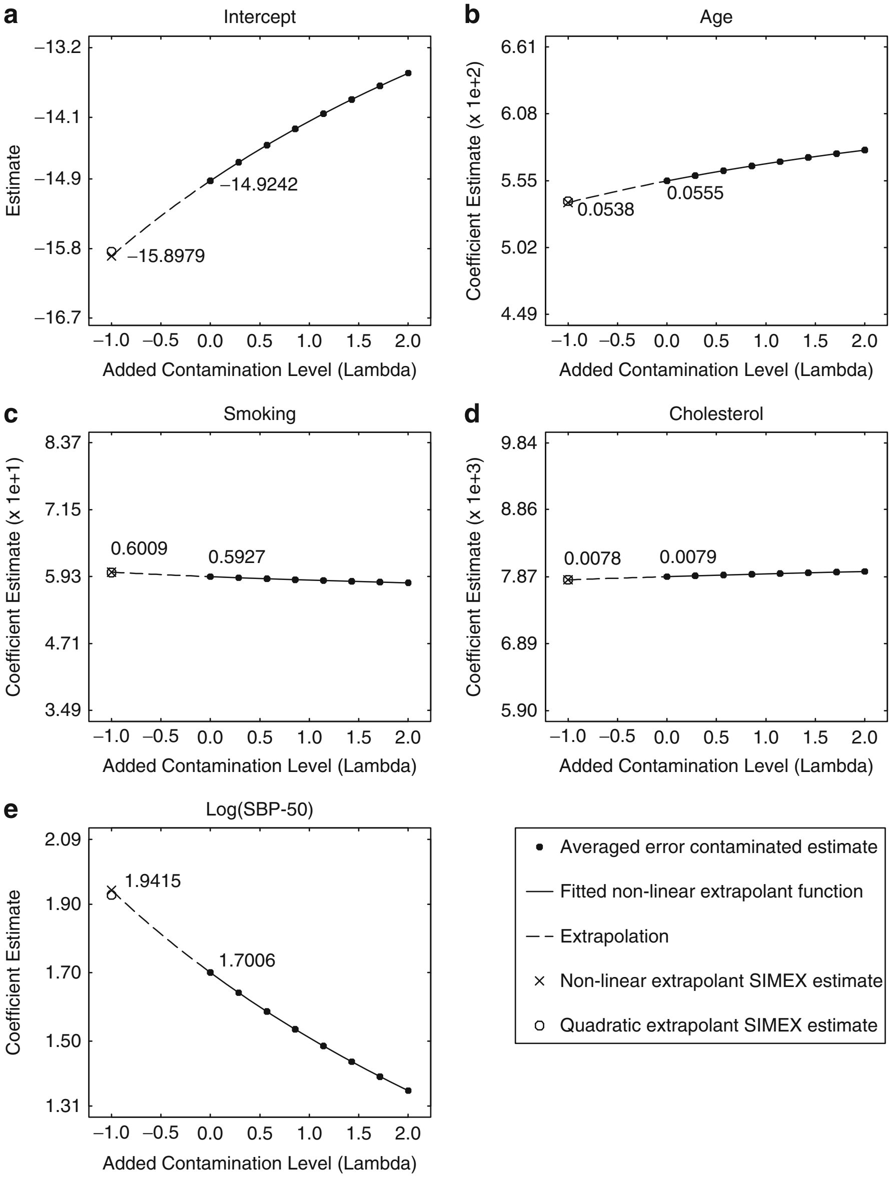


Fig. 33.1 SIMEX extrapolation plots for the Framingham data. Vertical axis scaling is ± two standard errors of the naive estimates

induces bias in estimating equations, which translates into bias in the parameter estimator. Modifying a score to remove bias produces estimators without bias.

To illustrate this concept, consider the no-intercept simple linear regression model with classical measurement error; $Y = \beta_x X + \epsilon$, $W = X + \sigma_u U$ and where X, ϵ, and U have mean zero. In the absence of measurement error, the least squares estimator for β_x solves $\sum_{i=1}^{n} \psi(Y_i, X_i; \beta_x) = 0$ where $\psi(Y_i, X_i; \beta_x) = (Y_i - \beta_x X_i)X_i$ is the least squares score. The score is unbiased: $E[\psi(Y_i, X_i; \beta_x)] = \beta_x \sigma_x^2 - \beta_x \sigma_x^2 = 0$. The score is no longer unbiased when W replaces X; $E[\psi(Y_i, W_i; \beta_x)] = \beta_x \sigma_x^2 - \beta_x(\sigma_x^2 + \sigma_u^2) \neq 0$ whenever $\sigma_u^2 > 0$ and $\beta_x \neq 0$.

Corrected scores are unbiased estimators of the score that would be used in the absence of measurement error. A corrected score $\psi^*(Y_i, W_i; \beta_x)$ satisfies $E[\psi^*(Y_i, W_i; \beta_x)] = \psi(Y_i, X_i; \beta_x)$ where the expectation is with respect to the measurement error distribution. Corrected scores were first defined in Stefanski (1989a) and Nakamura (1990). Note that corrected scores are unbiased whenever the original score is unbiased. This means that estimators obtained from corrected scores are fully consistent.

The corrected score for the simple linear no-intercept regression model is easily seen to be $\psi^*(Y_i, W_i; \beta_x) = \psi(Y_i, W_i; \beta_x) + \sigma_u^2 \beta_x$ resulting in the estimator $\widehat{\beta}_x = \sum_{i=1}^{n} Y_i W_i / (\sum_{i=1}^{n} W_i^2 - \sigma_u^2)$. In applications, an estimate of the measurement error variance replaces σ_u^2. Note that the corrected score estimator for the linear model is also the method-of-moments estimator.

For the linear model, the corrected score was identified without making an assumption on the distribution of the measurement error. For non-linear regression models, obtaining a corrected score generally requires specification of the measurement error distribution. Typically the normal distribution is assumed.

To illustrate the corrected score method for a non-linear model, consider Poisson regression with no intercept. The likelihood score in the absence of measurement error is $\psi(Y_i, X_i; \beta_x) = (Y_i - \exp\{\beta_x X_i\})X_i$. If we assume that the measurement error satisfies $U \sim N(0, 1)$, then $\psi^*(Y_i, W_i; \beta_x) = (Y_i - \exp\{\beta_x W_i - \beta_x^2 \sigma_u^2/2\})W_i + \beta_x \sigma_u^2 \exp\{\beta_x W_i - \beta_x^2 \sigma_u^2/2\})$ is the corrected score. Using results for the moment-generating function of a normal random variable, one can verify that $E[\psi^*(Y_i, W_i; \beta_x)] = (Y_i - \exp\{\beta_x X_i\})X_i$ where the expectation is with respect to the measurement error. The corrected score estimator $\widehat{\beta}_x$ solves $\sum_{i=1}^{n} \psi^*(Y_i, W_i; \widehat{\beta}_x) = 0$, and the solution must be obtained numerically for Poisson regression.

It is not always possible to obtain a corrected score (Stefanski 1989a). For example, the likelihood scores for logistic and probit regression do not admit a corrected score, except under certain restrictions (Buzas and Stefanski 1996c; Buzas 2009). Methods for obtaining corrected scores and approximately corrected scores for a general class of models via computer simulation have been recently studied (Novick and Stefanski 2002; Devanarayan and Stefanski 2002). Buonaccorsi (1996) describes another approach for correcting score functions in the presence of covariate measurement error that is equivalent to the corrected score approach for some disease models.

33.4.5 Conditional Scores

Conditional score estimation is the default method for logistic regression when the classical additive error model holds. The statistical theory of sufficient statistics and maximum likelihood underlie the derivation of conditional scores, and conditional score estimators retain certain optimality properties of likelihood estimators. Though we focus on logistic regression here, the method applies to a broader class of regression models, including Poisson and gamma regression. The method was derived in Stefanski and Carroll (1987). Construction of the conditional score estimator assumes that the measurement error is normally distributed. However, the estimator remains effective in reducing bias and is surprisingly efficient for modest departures from the normality assumption (Huang and Wang 2001). Derivation of conditional scores and consistency of estimators does not require an assumption on the distribution of X. Computing conditional score estimators requires an estimate of the measurement error variance.

The conditional score estimator for logistic regression is defined implicitly as the solution to estimating equations that are closely related to the maximum likelihood estimating equations used in the absence of measurement error. In the absence of measurement error, the maximum likelihood estimator of $(\beta_1, \beta_z, \beta_x)$ is defined implicitly as the solution to

$$\sum_{i=1}^{n} \{Y_i - F(\beta_1 + \beta_z Z_i + \beta_x X_i)\} \begin{pmatrix} 1 \\ Z_i \\ X_i \end{pmatrix} = 0,$$

where $F(v) = \{1 + \exp(-v)\}^{-1}$ is the logistic distribution function. The conditional score estimator is defined as the solution to the equations

$$\sum_{i=1}^{n} \{Y_i - F(\beta_1 + \beta_z Z_i + \beta_x \Delta_i)\} \begin{pmatrix} 1 \\ Z_i \\ \Delta_i \end{pmatrix} = 0$$

where $\Delta_i = W_i + (Y_i - \frac{1}{2})\widehat{\sigma}_u^2 \beta_x$ and $\widehat{\sigma}_u^2$ is an estimate of the measurement error variance. Conditional score estimation for logistic regression replaces the unobserved X_i with Δ_i. It can be shown that $E[Y \mid Z, \Delta] = F(\beta_1 + \beta_z Z + \beta_x \Delta)$, and it follows that the conditional score is unbiased. Because Δ depends on the parameter β_x, it is not possible to estimate $(\beta_1, \beta_z, \beta_x)$ using standard software by replacing X with Δ. Standard errors are computed using the sandwich estimator or bootstrap.

For models other than the logistic, the simple scheme of replacing X with Δ is not true generally, and conditional score estimating equations for Poisson and gamma regression is much more complicated.

The conditional score estimator for the logistic model compares favorably in terms of efficiency to the full maximum likelihood estimator that requires specification of an exposure model (Stefanski and Carroll 1990b). Sugar et al. (2007) derive the conditional score for the logistic model with a flexible exposure model.

Construction of a conditional score requires availability of a *sufficient statistic* for covariates measured with error. For some models, for example, logistic regression with a quadratic term in X, a sufficient statistic is not available and a conditional score does not exist. Tsiatis and Ma (2004) extend the conditional score approach by deriving estimating equations that retain certain optimality properties and are consistent regardless of the distribution for X. Their method is applicable to a broad class of disease models and essentially reproduces the conditional score method when there is a sufficient statistic. Computation of their estimator is non-trivial and the focus of current research.

33.4.6 Instrumental Variables

The methods described so far require additional data that allow estimation of the measurement error variance. Replicate observations and internal/external validation data are two sources of such additional information. Instrumental variables are another source of information that identify regression parameters in the presence of covariate measurement error. Instrumental variables (IV), denoted T, are additional measurements of X that satisfy three requirements; (i) T is non-differential, that is, $f_{Y|Z,X,T} = f_{Y|Z,X}$; (ii) T is correlated with X; and (iii) T is independent of $W - X$. Note that a replicate observation is an instrumental variable but an instrumental variable is not necessarily a replicate. It is possible to use an instrumental variable to estimate the measurement error variance and then use one of the above methods. Doing so can be inefficient, and IV methods typically do not directly estimate the measurement error variance.

Consider the cancer case-control study of arsenic exposure mentioned in Sect. 33.3. Two measurements of arsenic exposure are available for each case/control in the form of drinking water and toenail concentrations. Neither measure is an exact measure of long-term arsenic exposure (X). Taking toenail concentration to be an unbiased measurement of X, the drinking water concentration can serve as an instrumental variable.

Instrumental variable methods have been used in linear measurement error models since the 1940s, see Fuller (1987) for a good introduction. Instrumental variable methods for non-linear models were first studied in Amemiya (1985, 1990a, b). Extensions of regression calibration and conditional score methodology to instrumental variables are given in Carroll and Stefanski (1994), Stefanski and Buzas (1995), and Buzas and Stefanski (1996a, b). Hu and Schennach (2008) show that instrumental variables identify broad classes of regression models. In this regard, see also Carroll et al. (2004). Gustafson (2007) examines implications when the IV non-differential assumption (assumption (i) above) is mildly violated.

The essential idea underlying instrumental variable estimation can be understood by studying the simple linear model without intercept: $Y = \beta_x X + \epsilon$ and $W = X + \sigma_u U$. Then $Y = \beta_x W + \tilde{\epsilon}$ where $\tilde{\epsilon} = \epsilon - \beta_x \sigma_u U$, and it appears that Y and W follow a simple linear regression model. However, W and $\tilde{\epsilon}$ are correlated, violating a standard assumption in linear regression, and the least squares estimator for β_x is biased; see Sect. 33.2. The least squares estimating equation $\sum_{i=1}^{n}\{Y_i - \beta_x W_i\}W_i = 0$ is biased because W_i and $Y_i - \beta_x W_i$ are correlated. This suggests that an unbiased equation can be constructed by replacing W_i outside the brackets with a measurement uncorrelated with $Y_i - \beta_x W_i$. An IV T satisfies the requirement and the IV estimating equation $\sum_{i=1}^{n}\{Y_i - \beta_x W_i\}T_i = 0$ results in the consistent estimator $\widehat{\beta}_x = \sum_{i=1}^{n} Y_i T_i / \sum_{i=1}^{n} W_i T_i$. Non-zero correlation between X and T is required so that the denominator is not estimating zero. The key idea is that the score factors into two components where the first component $\{Y_i - \beta_x W_i\}$ has expectation zero and the second component T_i is uncorrelated with the first.

The method must be modified for non-linear problems. Logistic regression will be used to illustrate the modification. If we ignore measurement error, the estimating equations for logistic regression are

$$\sum_{i=1}^{n}\{Y_i - F(\beta_1 + \beta_z Z_i + \beta_x W_i)\}\begin{pmatrix}1\\ Z_i\\ W_i\end{pmatrix} = 0.$$

Unlike the linear case, for the logistic model and non-linear models generally, the first term in the estimating score, $\{Y_i - F(\beta_1 + \beta_z Z_i + \beta_x W_i)\}$, does not have expectation zero, so that replacing W_i with T_i outside the brackets in the above equation does not result in an estimator that reduces bias.

Define the logistic regression instrumental variable estimating equations

$$\sum_{i=1}^{n} h(Z_i, W_i, T_i)\{Y_i - F(\beta_1 + \beta_z Z_i + \beta_x W_i)\}\begin{pmatrix}1\\ Z_i\\ T_i\end{pmatrix} = 0$$

where $h(Z_i, W_i, T_i) = \sqrt{\frac{F'(\beta_1 + \beta_z Z_i + \beta_x T_i)}{F'(\beta_1 + \beta_z Z_i + \beta_x W_i)}}$ is a scalar-valued weight function and F' denotes the derivative of F. It can be shown the estimating equation is unbiased provided the distribution of the measurement error is symmetric, implying the estimator obtained from the equations is fully consistent. See Buzas (1997) for extensions to other disease models, including the Poisson and gamma models. Huang and Wang (2001) provide an alternative approach for the logistic model.

33.4.7 Likelihood Methods

Likelihood methods for estimation and inference are appealing because of optimality properties of maximum likelihood estimates and dependability of likelihood ratio confidence intervals. In the context of measurement error problems, the advantages of likelihood methods relative to functional methods have been studied in Schafer and Purdy (1996) and Küchenhoff and Carroll (1997). However, the advantageous properties are contingent on correct specification of the likelihood. As discussed below, this is often a difficult task in measurement error problems.

The likelihood for an observed data point (Y, W) conditional on Z is

$$f_{YW|Z} = \int f_{Y|Z,X,W} f_{W|Z,X} f_{X|Z} dx = \int f_{Y|Z,X} f_{W|Z,X} f_{X|Z} dx$$

where the second equality follows from the assumption of non-differential measurement error and W is possibly a vector containing replicates. The integral is replaced by a sum if X is a discrete random variable. The likelihood for the observed data is $\prod_{i=1}^{N} f_{Y_i,W_i|Z_i}$, and maximum likelihood estimates are obtained by maximizing the likelihood over all the unknown parameters in each of the three component distributions comprising the likelihood. In principle, the procedure is straightforward. However, there are several important points to be made.

1. The likelihood for the observed data requires *complete* distributional specification for the disease model $(f_{Y|Z,X})$, the error model $(f_{W|Z,X})$, and an exposure model $(f_{X|Z})$.
2. As was the case for functional models, estimation of parameters in the disease model generally requires, for all intents and purposes, observations that allow estimation of parameters in the error model, for example, replicate measurements.
3. When the exposure is modeled as a continuous random variable, for example, the normal distribution, the likelihood requires evaluation of an integral. For many applications, the integral cannot be evaluated analytically, and numerical methods must be used, typically Gaussian quadrature or Monte Carlo methods.
4. Finding the maximum of the likelihood is not always straightforward.

While the last two points must be addressed to implement the method, they are technical points and will not be discussed in detail. In principle, numerical integration followed by a maximization routine can be used, but this approach is often difficult to implement in practice; see Schafer (2002). Algorithms for computation and maximization of the likelihood in general regression models with exposure measurement error are given in Higdon and Schafer (2001) and Schafer (2002). Alternatively, a Bayesian formulation can be used to circumvent some of the computational difficulties; see Sect. 33.4.8. For the normal theory linear model and probit regression with normal distribution for the exposure model, the likelihood can be obtained analytically (Fuller 1987; Carroll et al. 1984; Schafer 1993). The analytical form of the likelihood for the probit model often provides an adequate approximation to the likelihood for the logistic model.

The first point above deserves discussion. None of the preceding methods required specification of an exposure model (functional methods). Here an exposure model is required. It is common to assume $X \mid Z \sim N(\alpha_1 + \alpha_x Z, \sigma^2_{x|z})$, but unless there are validation data, it is not possible to assess the adequacy of the exposure model using the data. Some models are robust to the normality assumption. For example, in the normal theory linear model, that is, when (Y, Z, X, W) is jointly normal, maximum likelihood estimators are fully consistent regardless of the distribution of X. For other disease models, the robustness to misspecification is not always clear-cut, but can be tested; see Huang et al. (2006) for a statistical test assessing the robustness of parameter estimates to the choice of distribution for X.

In a Bayesian framework, Richardson and Leblond (1997) show misspecification of the exposure model can affect estimation for logistic disease models.

Semi-parametric and flexible parametric modeling are two approaches that have been explored to address potential robustness issues in specifying an exposure model. Semi-parametric methods leave the exposure model unspecified, and the exposure model is essentially considered as another parameter that needs to be estimated. These models have the advantage of model robustness but may lack efficiency relative to the full likelihood, see Roeder et al. (1996), Schafer (2001), and Taupin (2001).

Flexible parametric exposure models typically use a mixture of normal densities to model the exposure distribution, as normal mixtures are capable of capturing moderately diversified features of distributions. Flexible parametric approaches have been studied in Küchenhoff and Carroll (1997), Carroll et al. (1999b), Schafer (2002), Gustafson et al. (2002), and Roy and Banerjee (2006). Guolo (2008) uses a skew-normal distribution to model the exposure in a case-control setting.

The likelihood can also be obtained conditional on both W and Z. In this case, the likelihood is

$$f_{Y|Z,W} = \int f_{Y|Z,X} f_{X|Z,W} dx$$

necessitating an exposure model relating X to W and Z. This form of the likelihood is natural for Berkson error models. In general, the choice of which likelihood to use is a matter of modeling convenience.

33.4.8 Bayesian Methods

Research and application of Bayesian methods (see also chapter ▶Bayesian Methods in Epidemiology of this handbook) in nearly all areas of statistics have greatly increased since the late 1980s with the advent of algorithms for sampling from posterior distributions. The field of measurement error is no exception.

Bayesian models are conceptually straightforward, and from an operational standpoint, Bayesian models can be considered an extension of likelihood models. A Bayesian model starts with the likelihood model, that is, disease, exposure, and error models, and adds a prior probability model to all unknown parameters in the likelihood. The posterior distribution for the unknown parameters is in principle obtained using Bayes theorem. The posterior distribution is the basis of inference in the Bayesian paradigm.

The Bayesian machinery can be applied to any data analytical setting where a likelihood is postulated, for example, covariate measurement error problems in multiple regression, logistic regression, survival analysis, non-parametric regression, and so on. While the Bayesian method is straightforward to describe and widely applicable, it is not necessarily straightforward to implement. Direct application of Bayes theorem typically requires evaluation of analytically and computationally intractable integrals.

Let $f_{Y|X,Z,\theta_Y}$, $f_{W|X,Z,\theta_W}$, and $f_{X|Z,\theta_X}$ denote the probability models for the outcome, error, and exposure where θ_Y, θ_W, and θ_X denote the unknown parameters in the respective models. A Bayesian analysis requires a prior probability distribution for the unknown parameters $\theta = (\theta_Y, \theta_W, \theta_X)$. Denote the prior density by f_θ. Usually the priors for θ_Y, θ_W, and θ_X are independent, so that $f_\theta = f_{\theta_Y} f_{\theta_W} f_{\theta_X}$.

Suppose the observable data consist of Y_i, W_{ij}, Z_i for $i = 1, \ldots n$ and $j = 1, \ldots, r_i$. The posterior distribution for the unobservable vector of covariates $X = (X_1, \ldots, X_n)$ and the model parameters θ is given by Bayes theorem:

$$f_{X,\theta|Y,W,Z} = \frac{\prod_{i=1}^{n} f_{Y_i|X_i,Z_i,\theta_Y} \prod_{j=1}^{r_i} f_{W_{ij}|X_i,Z_i,\theta_W} f_{X_i|Z_i,\theta} f_\theta}{\int\int \prod_{i=1}^{n} f_{Y_i|X_i,Z_i,\theta_Y} \prod_{j=1}^{r_i} f_{W_{ij}|X_i,Z_i,\theta_W} f_{X_i|Z_i,\theta} f_\theta dx d\theta}.$$

The integral in the denominator of the right-hand side is typically analytically intractable and of high dimension, with the latter issue rendering numerical integration techniques ineffective. Furthermore, we are interested in the marginal posterior distribution of θ, requiring another integration that is typically difficult to evaluate: $f_{\theta|Y,W,Z} = \int f_{X,\theta|Y,W,Z} dx$.

Evaluation of these integrals is circumvented using Markov chain Monte Carlo (MCMC) algorithms, with the Gibbs sampler and the Metropolis-Hastings algorithms, the most widely used variants (see chapter ▸Bayesian Methods in Epidemiology of this handbook). MCMC algorithms are able to sample from the posterior distribution $f_{X,\theta|Y,W,Z}$ without evaluating the integrals on the right-hand side. The marginal distribution $f_{\theta|Y,W,Z}$ can be estimated using only the sampled θ's. A detailed description of these algorithms is beyond the scope of this chapter. Carroll et al. (2006) and Gustafson (2004) provide details of implementing MCMC methods for covariate measurement error problems.

An important consideration when using MCMC methods is assessing convergence and mixing of the algorithm. Trace plots are often the best diagnostic tool for assessing convergence and mixing, though other diagnostic procedures have been developed. It is often helpful to compute several "chains," each starting in a different region of the parameter space, and noting whether the different chains end up in approximately the same region of the posterior distribution. Convergence and mixing of the separate chains are assessed visually using trace plots and quantitatively using Gelman and Rubin's R statistic. The example below illustrates the use of trace plots.

To illustrate the specification of commonly employed uninformative priors in a measurement error setting, consider the normal theory linear regression model:

$$\begin{aligned} Y \mid X, \theta_Y &\sim N(\beta_0 + \beta_x X, \sigma^2) \\ W \mid X, \theta_W &\sim N(X, \sigma_u^2) \\ X \mid \theta_X &\sim N(\mu_x, \sigma_x^2), \end{aligned}$$

where the distributions are independent. For this model, $\theta_Y = (\beta_0, \beta_x, \sigma^2)$, $\theta_W = \sigma_u^2$ and $\theta_X = (\mu_x, \sigma_x^2)$. Using independent normal prior distributions with mean zero and large variance for β_0, β_x, and μ_x, and independent inverse gamma distributions for the variance components σ^2, σ_u^2, and σ_x^2 allows for straightforward implementation of the Gibbs sampler, as these priors are conjugate for the normal theory model above. Using large variances for the normal distributions and appropriate shape and scale parameters for the inverse gamma distribution results in uninformative priors. Other choices of priors can also be used. For example, diffuse uniform priors for $(\beta_0, \beta_x, \mu_x, \sigma, \sigma_u, \sigma_x)$ are also commonly employed.

For most models, covariate measurement error renders some model parameters non-identified; see the discussion in Sect. 33.2.3. Additional data, such as replicate covariate measurements, instrumental variables, or validation data, can be used to identify model parameters. In the Bayesian framework, posterior distributions for non-identified parameters using proper priors still obtain. However, the posterior distributions of non-identified parameters will not converge to the true parameter values. Nevertheless, Gustafson (2005, 2009) recently studied the amount of information available for parameters in non-identified models and shows that in certain scenarios a moderate amount of prior information can lead to inferences that outperform inferences from identified models obtained by contracting (or expanding) the non-identified model.

33.4.8.1 Example

We return again to the Framingham data described in Sect. 33.2.6. Measurements of coronary heart disease, age, and three measurements of the logarithm of systolic blood pressure, denoted by Y, Z, W_1, W_2, and W_3, respectively, were obtained for $n = 1,615$ individuals. The log systolic blood pressure measurements are assumed to be replicates with additive normal measurement error, that is, $W_i = X_i + \sigma_u U_i, i = 1, 2, 3$ where $U_i \sim N(0, 1)$ are independent. For the Bayes fit, we assume $X \sim N(\mu_x, \sigma_x^2)$. Interest is in fitting the logistic regression model:

$$\text{logit}(E[Y \mid X, Z]) = \beta_0 + \beta_x X + \beta_z Z.$$

Having replicate measurements identifies all model parameters.

Four estimators were computed for comparison: the naive estimator that uses W_1 and Z as covariates, the Bayes, conditional score, and corrected score estimators (Buzas 2009). The Bayesian estimator used diffuse independent normal priors for each of β_0, β_x, β_z, and μ_x. Diffuse independent uniform priors were used for σ_x and σ_u. Three chains, each of length 10,000, were computed. The first 5,000 iterations were discarded, and the remaining samples thinned resulting in a sample of 1,000 from the posterior distribution. Convergence and mixing were very good; see the trace plots in Fig. 33.2. The plots show a stable mean and variance of the posterior samples and rapid movement through the posterior distribution. It is worth noting that if age and log systolic blood pressure were not centered and scaled during fitting, mixing was very slow for β_0 and β_x, requiring several hundred thousand iterations of the chain. Computations were done using WinBUGS software.

As seen in Table 33.3, the posterior mean and standard deviation from the Bayesian analysis are very similar to the estimates obtained from the conditional and corrected scores.

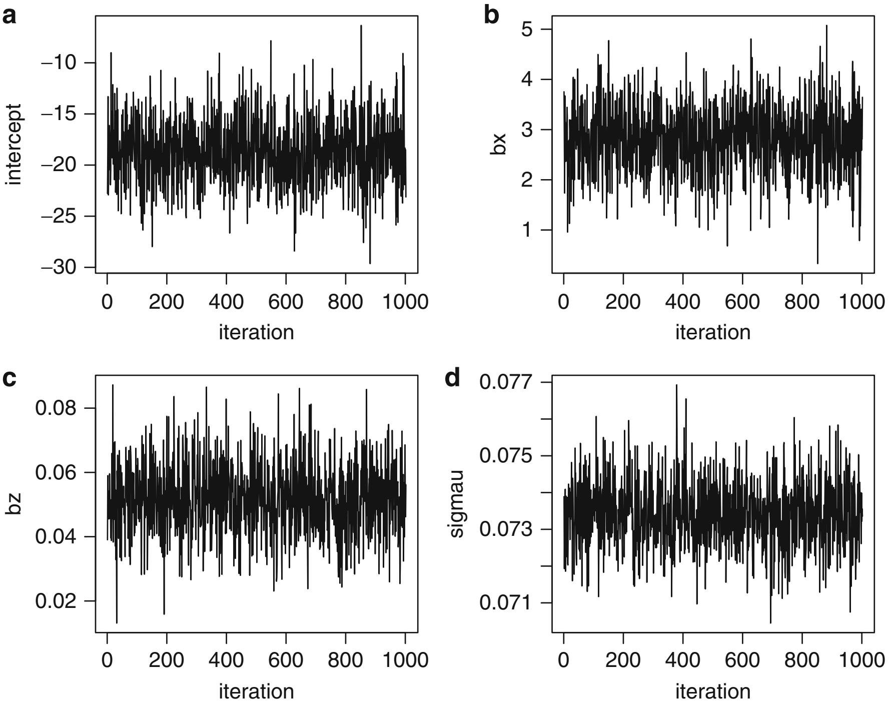


Fig. 33.2 Trace plots for the Framingham data

Table 33.3 Logistic parameter estimates for Framingham data

Covariate	Naive	Bayes	Conditional score	Corrected score
Intercept	−15.565 (2.624)	−18.655 (3.287)	−18.603 (3.288)	−18.614 (3.350)
Log blood pressure	2.161 (0.547)	2.806 (0.690)	2.817 (0.696)	2.835 (0.714)
Age	0.053 (0.010)	0.052 (0.011)	0.049 (0.011)	0.048 (0.010)

To summarize, advantages of Bayesian analyses include wide applicability, straightforward inference using posterior distributions, and the potential to handle complex and perhaps non-identified models. The disadvantages include potential sensitivity to priors and exposure models, the potential need to develop application-specific MCMC algorithms, and assessment of convergence of MCMC samplers. Note that more model complexity requires more checks on sensitivity of model assumptions.

33.4.8.2 Additional Notes

WinBUGS is currently the most widely used software package that implements MCMC algorithms. WinBUGS has a great deal of flexibility, and many types of outcome, exposure, error models, and priors can be specified. The syntax for WinBUGS is similar to that of R, though not identical. A strength of WinBUGS is that it is relatively simple to specify hierarchical Bayesian models using syntax analogous to mathematical notation one would use to specify these models on paper. Currently WinBUGS is free.

Perhaps the earliest paper to study covariate measurement error in epidemiology from a Bayesian perspective is Richardson and Gilks (1993). Examples of recent research into Bayesian methods for covariate measurement error in epidemiological settings are the following: Johnson et al. (2007) develop models in the setting of nutritional epidemiology for studying the effects of folate intake, measured imprecisely, of expectant mothers on gestational age. Sinha et al. (2010) provide a Bayesian analysis using semi-parametric models for both the disease and exposure models. Cheng and Crainiceanu (2009) develop an MCMC algorithm for survival data using non-parametric models for the baseline hazard and survival-exposure relation. They apply their method to the study of chronic kidney disease using an imprecise measure of baseline kidney function as a covariate. Li et al. (2007) use Bayesian methods in the context of estimating the effects of radiation exposure on thyroid disease. The problems in these papers share the feature that the models studied are complex, a quality Bayesian methods are often adept at handling.

33.4.9 Survival Analysis

The preceding discussion described general strategies applicable to a broad range of disease models. Survival data warrants its own section, as analysis of survival data with exposure measurement error using proportional hazards models presents some new issues (see also chapter ▶Exposure Assessment of this handbook). Of the methods presented, only SIMEX and Bayesian methods (when a likelihood is specified) can be applied without modification in the proportional hazards setting.

Many of the proposed methods for measurement error correction in proportional hazards models fall into one of two general strategies. The first strategy is to approximate the induced hazard and then use the approximated hazard in the partial likelihood equations. This strategy is analogous to the regression calibration approximation discussed earlier. The second strategy is to modify the partial likelihood estimating equations. Methods based on this strategy stem from the corrected and conditional score paradigms.

In the absence of measurement error, the proportional hazards model postulates a hazard function of the form $\lambda(t \mid Z, X) = \lambda_0(t) \exp(\beta_z^T Z + \beta_x X)$ where $\lambda_0(t)$ is an unspecified baseline hazard function. Estimation and inference for (β_x, β_z) are carried out through the partial likelihood function, as it does not depend on $\lambda_0(t)$.

Prentice (1982) has shown that when (Z, W) is observed, the induced hazard is $\lambda(t \mid Z, W) = \lambda_0(t) E[\exp(\beta_z^T Z + \beta_x X) \mid T \geq t, Z, W]$. The induced hazard requires a model for X conditional on $(T \geq t, Z, W)$. This is problematic because the distribution of T is left unspecified in proportional hazards models. However, when the disease is rare $\lambda(t \mid Z, W) \approx \lambda_0(t) E[\exp(\beta_z^T Z + \beta_x X) \mid Z, W]$ (Prentice 1982) and if we further assume that $X \mid Z, W$ is approximately normal with constant variance, then the induced hazard is proportional to

$\exp(\beta_z^T Z + \beta_x E[X \mid Z, W])$. In other words, regression calibration is appropriate in the proportional hazards setting when the disease is rare and $X \mid Z, W$ is approximately normal.

Modifications to the regression calibration algorithm have been developed for applications where the rare disease assumption is untenable; see Clayton (1991), Tsiatis et al. (1995), Wang et al. (1997) and Xie et al. (2001). Conditioning on $T \geq t$ cannot be ignored when the disease is not rare. The idea is to reestimate the calibration function $E[X \mid Z, W]$ in each risk set, that is, the set of individuals known to be at risk at time t. Clayton's proposal assumes that the calibration functions across risk sets have a common slope and his method can be applied provided one has an estimate of the measurement error variance. Xie et al. (2001) extend the idea to varying slopes across the risk sets and require replication (reliability data). Tsiatis et al. (1995) consider time-varying covariates and also allow for varying slopes across the risk sets.

When a validation subsample is available, it is possible to estimate the induced hazard non-parametrically, that is, without specifying a distribution for $X \mid (T \geq t, Z, W)$; see Zhou and Pepe (1995) and Zhou and Wang (2000) for the cases when the exposure is discrete and continuous, respectively.

The second strategy mentioned above avoids modeling the induced hazard and instead modifies the partial likelihood estimating equations. Methods based on the corrected score concept are explored in Nakamura (1992), Buzas (1998), Huang and Wang (2000), Augustin (2004), and Yi and Lawless (2007). The methods in Nakamura (1992) and Buzas (1998) assume the measurement error is normally distributed and only require an estimate of the measurement error variance. In contrast, the approach in Huang and Wang (2000) does not require assumptions on the measurement error distribution, but replicate observations on the mismeasured exposure are needed to compute the estimator. Each of the methods has been shown to be effective in reducing bias in parameter estimators. Tsiatis and Davidian (2001) extend conditional score methodology to the proportional hazards setting with covariates possibly time dependent. Song and Huang (2005) compare the conditional and corrected score approaches in small samples, generally noting the conditional score approach is superior.

33.5 Conclusions

Epidemiologists have long recognized the importance of addressing problems of measurement and ascertainment in the statistical analysis of epidemiological data. Much of the research in measurement error models has its origins in specific epidemiological applications, and this is reflected by many of the research papers cited in this review article. The importance of measurement error modeling to epidemiological research is on the rise, and that trend is likely to continue for the foreseeable future.

As the understanding of the etiology of disease increases, so too will the sophistication of the statistical models used to extract information from epidemiological data. Success in these modeling endeavors will depend on the ability to accurately model ever finer sources of variability in data, and measurement error is frequently one such non-negligible source of variation. Recent articles studying epidemiological applications with covariate measurement error bare out the increasing complexity of models studied.

This chapter provides an introduction to and a review of the literature on the problem of statistical inference in the presence of measurement error. Section 33.2 discussed the effects of measurement error in common epidemiological models. With the inevitable increase in the sophistication of models for epidemiological data, there will be a need to understand the effects of measurement error on parameter estimation in biologically based and physiologically based models of disease. The increasing ability to collect very detailed and precise information on subjects in validation samples (e.g., via continuous monitoring of biological processes, or genetic screening) means that consideration of measurement error at the design stage of a study will take on greater importance. Hence, the timely relevance of the issues and methods is discussed in Sect. 33.3. More elaborate modeling places greater demands on methods of estimation. Section 33.4 provided a summary and review of common approaches to estimation in the presence of measurement. Future research will necessarily have to accommodate more complex models and possibly multiple variates measured with correlated error.

References

Amemiya Y (1985) Instrumental variable estimator for the nonlinear errors in variables model. J Econom 28:273–289

Amemiya Y (1990a) Instrumental variable estimation of the nonlinear measurement error model. In: Brown PJ, Fuller WA (eds) Statistical analysis of measurement error models and application. American Mathematics Society, Providence

Amemiya Y (1990b) Two-stage instrumental variable estimators for the nonlinear errors in variables model. J Econom 44:311–332

Armstrong BK, White E, Saracci R (1992) Principles of exposure measurement error in epidemiology. Oxford University Press, Oxford

Augustin T (2004) An exact corrected log-likelihood function for Cox's proportional hazards model under measurement error and some extensions. Scand J Stat 31:43–50

Berkson J (1950) Are there two regressions? J Am Stat Assoc 45:164–180

Buonaccorsi JP (1988) Errors in variables with systematic biases. Commun Stat Theory Methods 18:1001–1021

Buonaccorsi JP (1996) A modified estimating equation approach to correcting for measurement error in regression. Biometrika 83:433–440

Buonaccorsi JP (2010) Measurement error: models, methods, and applications. Chapman and Hall/CRC, London

Buzas JS (1997) Instrumental variable estimation in nonlinear measurement error models. Commun Stat Theory Method 26:2861–2877

Buzas JS (1998) Unbiased scores in proportional hazards regression with covariate measurement error. J Stat Plan Inference 67:247–257

Buzas JS (2009) A note on corrected scores for logistic regression. Stat Probab Lett 79:2351–2358

Buzas JS, Stefanski LA (1996a) Instrumental variable estimation in probit measurement error models. J Stat Plan Inference 55:47–62

Buzas JS, Stefanski LA (1996b) Instrumental variable estimation in generalized linear measurement error models. J Am Stat Assoc 91:999–1006

Buzas JS, Stefanski LA (1996c) A note on corrected score estimation. Stat Probab Lett 28:1–8

Cain KC, Breslow NE (1988) Logistic regression analysis and efficient design for two-stage studies. Am J Epidemiol 128:1198–1206

Carroll RJ (1989) Covariance analysis in generalized linear measurement error models. Stat Med 8:1075–1093

Carroll RJ (1998) Measurement error in epidemiologic studies. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics. Wiley, New York, pp 2491–2519

Carroll RJ, Stefanski LA (1990) Approximate quasilikelihood estimation in models with surrogate predictors. J Am Stat Assoc 85:652–663

Carroll RJ, Stefanski LA (1994) Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. Stat Med 13:1265–1282

Carroll RJ, Spiegelman C, Lan KK, Bailey KT, Abbott RD (1984) On errors-in-variables for binary regression models. Biometrika 71:19–26

Carroll RJ, Gallo PP, Gleser LJ (1985) Comparison of least squares and errors-in-variables regression, with special reference to randomized analysis of covariance. J Am Stat Assoc 80:929–932

Carroll RJ, Küchenhoff H, Lombard F, Stefanski LA (1996) Asymptotics for the SIMEX estimator in structural measurement error models. J Am Stat Assoc 91:242–250

Carroll RJ, Maca JD, Ruppert D (1999a) Nonparametric regression in the presence of measurement error. Biometrika 86:541–554

Carroll RJ, Roeder K, Wasserman L (1999b) Flexible parametric measurement error models. Biometrics 55:44–54

Carroll RJ, Ruppert D, Crainiceanu CM, Tosteson TD, Karagas MR (2004) Nonlinear and nonparametric regression and instrumental variables. J Am Stat Assoc 99:736–750

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) Measurement error in nonlinear models, 2nd edn. Chapman and Hall, London

Cheng YJ, Crainiceanu CM (2009) Cox models with smooth functional effect of covariates measured with error. J Am Stat Assoc 104:11441154

Clayton DG (1991) Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In: Dwyer JH, Feinleib M, Lipsert P et al (eds) Statistical models for longitudinal studies of health. Oxford University Press, New York, pp 301–331

Cochran WG (1968) Errors of measurement in statistics. Technometrics 10:637–666

Cook J, Stefanski LA (1995) A simulation extrapolation method for parametric measurement error models. J Am Stat Assoc 89:1314–1328

de Gramont A, Buyse M, Abrahantes JC, Burzykowski T, Quinaux E, Cervantes A, Figer A, Lledo G, Flesch M, Mineur L, Carola E, Etienne PL, Rivera F, Chirivella I, Perez-Staub N, Louvet C, Andr T, Tabah-Fisch I, Tournigand C (2007) Reintroduction of oxaliplatin is associated with improved survival in advanced colorectal cancer. J Clin Oncol 25:3224–3229

Delaigle A, Hall P (2008) Using SIMEX for smoothing-parameter choice in errors-in-variables problems. J Am Stat Assoc 103:280–287

Devanarayan V, Stefanski LA (2002) Empirical simulation extrapolation for measurement error models with replicate measurements. Stat Probab Lett 59:219–225

Fearn T, Hill DC, Darby SC (2008) Measurement error in the explanatory variable of a binary regression: regression calibration and integrated conditional likelihood in studies of residential radon and lung cancer. Stat Med 27:2159–2176

Fleiss JL (1981) Statistical methods for rates and proportions. Wiley, New York

Fuller WA (1987) Measurement error models. Wiley, New York

Guolo A (2008) A flexible approach to measurement error correction in casecontrol studies. Biometrics 64:1207–1214

Gustafson P (2004) Measurement error and misclassification in statistics and epidemiology: impacts and bayesian adjustments. Chapman and Hall/CRC, London

Gustafson P (2005) On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. Stat Sci 20:111–140

Gustafson P (2007) Measurement error modelling with an approximate instrumental variable. J R Stat Soc Ser B Stat Methodol 69:797–815

Gustafson P (2009) What are the limits of posterior distributions arising from nonidentified models, and why should we care? J Am Stat Assoc 488:1682–1695

Gustafson P, Le ND, Valle M (2002) A Bayesian approach to case-control studies with errors in covariables. Biostatistics 3:229–243

Higdon R, Schafer DW (2001) Maximum likelihood computations for regression with measurement error. Comput Stat Data Anal 35:283–299

Holcomb JP (1999) Regression with covariates and outcome calculated from a common set of variables measured with error: estimation using the SIMEX method. Stat Med 18:2847–2862

Holcroft CA, Rotnitzky A, Robins JM (1997) Efficient estimation of regression parameters from multistage studies with validation of outcome and covariates. J Stat Plan Inference 65:349–374

Hu Y, Schennach SM (2008) Instrumental variable treatment of nonclassical measurement error models. Econometrica 76:195–216

Huang X, Stefanski LA, Davidian M (2006) Latent-model robustness in structural measurement error models. Biometrika 93:5364

Huang Y, Wang CY (2000) Cox regression with accurate covariates unascertainable: a nonparametric-correction approach. J Am Stat Assoc 95:1209–1219

Huang Y, Wang CY (2001) Consistent functional methods for logistic regression with errors in covariates. J Am Stat Assoc 95:1209–1219

Hughes MD (1993) Regression dilution in the proportional hazards model. Biometrics 49:1056–1066

Hwang JT, Stefanski LA (1994) Monotonicity of regression functions in structural measurement error models. Stat Probab Lett 20:113–116

IJmker S, Blatter BM, van der Beek AJ, van Mechelen W, Bongers PM (2006) Prospective research on musculoskeletal disorders in office workers (PROMO): study protocol. BMC Musculoskelet Disord 7:1–9

Johnson BA, Herring AH, Ibrahim JG, Siega-Riz AM (2007) Structured measurement error in nutritional epidemiology: applications in the pregnancy, infection, and nutrition (PIN) Study. J Am Stat Assoc 102:856–866

Karagas MR, Tosteson TD, Blum J, Morris SJ, Baron JA, Klaue B (1998) Design of an epidemiologic study of drinking water arsenic and skin and bladder cancer risk in a U.S. population. Environ Health Perspect 106:1047–1050

Karagas MR, Stukel TA, Tosteson TD (2002) Assessment of cancer risk and environmental levels of arsenic in New Hampshire. Int J Hyg Environ Health 205:85–94

Kim J, Gleser LJ (2000) SIMEX approaches to measurement error in ROC studies. Commun Stat Theory Method 29:2473–2491

Kim C, Hong C, Jeong M (2000) Simulation-extrapolation via the Bezier curve in measurement error models. Commun Stat Simul Comput 29:1135–1147

Kipnis V, Carroll RJ, Freedman LS, Li L (1999) Implications of a new dietary measurement error model for estimation of relative risk: application to four calibration studies. Am J Epidemiol 50:642–651

Kopecky KJ, Stepanenko V, Rivkind N, Voillequ P, Onstad L, Shakhtarin V, Parshkov E, Kulikov S, Lushnikov E, Abrosimov A, Troshin V, Romanova G, Doroschenko V, Proshin A, Tsyb A, Davis S (2006) Childhood thyroid cancer, radiation dose from chernobyl, and dose uncertainties in bryansk oblast, Russia: a population-based case-control study. Radiat Res 166:367–374

Kottgen A, Russell SD, Loehr LR, Crainiceanu CM, Rosamond WD, Chang PP, Chambless LE, Coresh J (2007) Reduced kidney function as a risk factor for incident heart failure: the atherosclerosis risk in communities (ARIC) study. J Am Soc Nephrol 18:1307–1315

Küchenhoff H, Carroll RJ (1997) Segmented regression with errors in predictors: semi-parametric and parametric methods. Stat Med 16:169–188

Küchenhoff H, Mwalili SM, Lesaffre E (2006) A general method for dealing with misclassification in regression: the misclassification simex. Biometrics 62:85–96

Lin X, Carroll RJ (1999) SIMEX variance component tests in generalized linear mixed measurement error models. Biometrics 55:613–619

Li Y, Lin X (2003) Functional inference in frailty measurement error models for clustered survival data using the SIMEX approach. J Am Stat Assoc 98:191–203

Li Y, Guolo A, Hoffman FO, Carroll RJ (2007) Shared uncertainty in measurement error problems, with application to nevada test site fallout data. Biometrics 63:1226–1236

MacMahon S, Peto R, Cutler J, Collins R, Sorlie P, Neaton J, Abbott R, Godwin J, Dyer A, Stamler J (1990) Blood pressure, stroke and coronary heart disease: part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. Lancet 335:765–774

Marcus AH, Elias RW (1998) Some useful statistical methods for model validation. Environ Health Perspect 106:1541–1550

McKeown-Eyssen GE, Tibshirani R (1994) Implications of measurement error in exposure for the sample sizes of case-control studies. Am J Epidemiol 139:415–421

Nakamura T (1990) Corrected score functions for errors-in-variables models: methodology and application to generalized linear models. Biometrika 77:127–137

Nakamura T (1992) Proportional hazards models with covariates subject to measurement error. Biometrics 48:829–838

Novick SJ, Stefanski LA (2002) Corrected score estimation via complex variable simulation extrapolation. J Am Stat Assoc 458:472–481

Prentice RL (1982) Covariate measurement errors and parameter estimation in a failure time regression model. Biometrika 69:331–342

Reilly M (1996) Optimal sampling strategies for two phase studies. Am J Epidemiol 143:92–100

Richardson S, Gilks WR (1993) A bayesian approach to measurement error problems in epidemiology using conditional independence models. Am J Epidemiol 138:430–442

Richardson S, Leblond L (1997) Some comments on misspecification of priors in Bayesian modelling of measurement error problems. Stat Med 16:203–213

Roeder K, Carroll RJ, Lindsay BG (1996) A nonparametric mixture approach to case-control studies with errors in covariables. J Am Stat Assoc 91:722–732

Rosner B, Spiegelman D, Willett WC (1990) Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Am J Epidemiol 132:734–745

Rosner B, Willett WC, Spiegelman D (1989) Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med 8:1051–1070

Roy S, Banerjee T (2006) A flexible model for generalized linear regression with measurement error. Ann Inst Stat Math 58:153–169

Schafer D (1993) Likelihood analysis for probit regression with measurement errors. Biometrika 80:899–904

Schafer D (2001) Semiparametric maximum likelihood for measurement error model regression. Biometrics 57:53–61

Schafer D (2002) Likelihood analysis and flexible structural modeling for measurement error model regression. J Comput Stat Data anal 72:33–45

Schafer D, Purdy K (1996) Likelihood analysis for errors-in-variables regression with replicate measurements. Biometrika 83:813–824

Sinha S, Mallick B, Kipnis V, Carroll RJ (2010) Semiparametric bayesian analysis of nutritional epidemiology data in the presence of measurement error. Biometrics 66:444–454

Song X, Huang Y (2005) On corrected score approach for proportional hazards model with covariate measurement error. Biometrics 61:702–714

Spiegelman D, Gray R (1991) Cost-efficient study designs for binary response data with Gaussian covariate measurement error. Biometrics 47:851–869

Spiegelman D, Carroll RJ, Kipnis V (2001) Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. Stat Med 20:139–160

Staudenmayer J, Ruppert D (2004) Local polynomial regression and simulation-extrapolation. J R Stat Soc Ser B 66:17–30

Stefanski LA (1985) The effects of measurement error on parameter estimation. Biometrika 72:583–592

Stefanski LA (1989) Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. Commun Stat Theory Method 18:4335–4358

Stefanski LA, Bay JM (1996) Simulation extrapolation deconvolution of finite population cumulative distribution function estimators. Biometrika 83:407–417

Stefanski LA, Buzas JS (1995) Instrumental variable estimation in binary measurement error models. J Am Stat Assoc 90:541–550

Stefanski LA, Carroll RJ (1985) Covariate measurement error in logistic regression. Ann Stat 13:1335–1351

Stefanski LA, Carroll RJ (1987) Conditional scores and optimal scores in generalized linear measurement error models. Biometrika 74:703–716

Stefanski LA, Carroll RJ (1990a) Score tests in generalized linear measurement error models. J R Stat Soc B 52:345–359

Stefanski LA, Carroll RJ (1990b) Structural logistic regression measurement error models. In: Brown PJ, Fuller WA (eds) Proceedings of the conference on measurement error models. Wiley, New York, pp 115–127

Stefanski LA, Cook J (1995) Simulation extrapolation: the measurement error jackknife. J Am Stat Assoc 90:1247–1256

Stram DO, Longnecker MP, Shames L, Kolonel LN, Wilkens LR, Pike MC, Henderson BE (1995) Cost-efficient design of a diet validation-study. Am J Epidemiol 142(3):353–362

Sugar EA, Wang C-Y, Prentice RL (2007) Logistic regression with exposure biomarkers and flexible measurement error. Biometrics 63:143–151

Taupin M (2001) Semi-parametric estimation in the nonlinear structural errors-in-variables model. Ann Stat 29:66–93

Thiebaut ACM, Freedman LS, Carroll RJ, Kipnis V (2007) Is it necessary to correct for measurement error in nutritional epidemiology? Ann Intern Med 146:65–67

Thomas D, Stram D, Dwyer J (1993) Exposure measurement error: influence on exposure-disease relationships and methods of correction. Ann Rev Public Health 14:69–93

Thürigen D, Spiegelman D, Blettner M, Heuer C, Brenner H (2000) Measurement error correction using validation data: a review of methods and their applicability in case-control studies. Stat Methods Med Res 9(5):447–74

Tosteson TD, Tsiatis AA (1988) The asymptotic relative efficiency of score tests in the generalized linear model with surrogate covariates. Biometrika 75:507–514

Tosteson TD, Ware JH (1990) Designing a logistic regression study using surrogate measures of exposure and outcome. Biometrika 77:11–20

Tosteson T, Stefanski LA, Schafer DW (1989) A measurement error model for binary and ordinal regression. Stat Med 8:1139–1147

Tosteson TD, Titus-Ernstoff L, Baron JA, Karagas MR (1994) A two-stage validation study for determining sensitivity and specificity. Environ Health Perspect 102:11–14

Tosteson TD, Buzas JS, Demidenko D, Karagas MR (2003) Power and sample size calculations for generalized regression models with covariate measurement error. Stat Med 22:1069–1082

Tsiatis AA, Davidian M (2001) A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. Biometrika 88:447–458

Tsiatis AA, Ma Y (2004) Locally efficient semiparametric estimators for functional measurement error models. Biometrika 91:835–848

Tsiatis AA, DeGruttola V, Wulfsohn MS (1995) Modeling the relationship of survival to longitudinal data measured with error: applications to survival and CD4 counts in patients with AIDS. J Am Stat Assoc 90:27–37

Wang CY, Hsu ZD, Feng ZD, Prentice RL (1997) Regression calibration in failure time regression. Biometrics 53:131–145

Wang N, Lin X, Gutierrez R, Carroll RJ (1998) Bias analysis and the SIMEX approach in generalized linear mixed effects models. J Am Stat Assoc 93:249–261

White E, Kushi LH, Pepe MS (1994) The effect of exposure variance and exposure measurement error on study sample size. Implications for design of epidemiologic studies. J Clin Epidemiol 47:873–880

Xie SX, Wang CY, Prentice RL (2001) A risk set calibration method for failure time regression by using a covariate reliability sample. J R Stat Soc B 63:855–870

Yi GY, Lawless JF (2007) A corrected likelihood method for the proportional hazards model with covariates subject to measurement error. J Stat Plan Inference 137:1816–1828

Zhou H, Pepe MS (1995) Auxiliary covariate data in failure time regression analysis. Biometrika 82:139–149

Zhou H, Wang CY (2000) Failure time regression with continuous covariates measured with error. J R Stat Soc Ser B 62:657–665

Missing Data

34

Geert Molenberghs, Caroline Beunckens, Ivy Jansen, Herbert Thijs, Geert Verbeke, and Michael G. Kenward

Contents

G. Molenberghs (✉) • H. Thijs • G. Verbeke
Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), University of Hasselt and Catholic University Leuven (KU Leuven), Diepenbeek, Belgium

C. Beunckens
BELGACOM, Brussels, Belgium

I. Jansen
Research Institute for Nature and Forest (INBO), Brussels, Belgium

M.G. Kenward
Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_20,

34.1 Introduction

The problem of dealing with missing values is common throughout statistical work and is present whenever human subjects are enrolled. Respondents may refuse participation or may be unreachable. Patients in clinical and epidemiological studies may withdraw their initial consent without further explanation. Early work on missing values was largely concerned with algorithmic and computational solutions to the induced lack of balance or deviations from the intended study design (Afifi and Elashoff 1966; Hartley and Hocking 1971). More recently, general algorithms such as the Expectation–Maximization (EM) (Dempster et al. 1977) and data imputation and augmentation procedures (Rubin 1987; Tanner and Wong 1987), combined with powerful computing resources, have largely provided a solution to this aspect of the problem. There remains the very difficult and important question of assessing the impact of missing data on subsequent statistical inference. Conditions can be formulated, under which an analysis that proceeds as if the missing data are missing by design, that is, ignoring the missing value process, can provide valid answers to study questions. While such an approach is attractive from a pragmatic point of view, the difficulty is that such conditions can rarely be assumed to hold with full certainty. Indeed, assumptions will be required that cannot be assessed from the data under analysis. Hence in this setting there cannot be anything that could be termed a definitive analysis, and hence any analysis of preference is ideally to be supplemented with a so-called sensitivity analysis.

In Sect. 34.2, the key illustrative case study is introduced. Section 34.3 is devoted to an overview of the fundamental concepts of incomplete longitudinal or multivariate data, including a general framework within which missing data ideas can be developed as well as a modeling framework for both Gaussian and non-Gaussian outcomes, useful to further develop and illustrate practical strategies to deal with incomplete data. Simple methods that are often used but carry quite a bit of danger in them are discussed in Sect. 34.4. In Sect. 34.5, an overview of methods is provided which are valid when missingness is missing at random (MAR), that is, when the probability of dropout is allowed to depend on the observed measurements. Such methods can be used as a primary analysis. In this chapter, the focus will be primarily on incomplete outcomes. However, incomplete covariate data, perhaps in conjunction with incomplete outcome data, are relatively common too. It should be noted, therefore, that multiple imputation, discussed in Sect. 34.5.3, can nicely handle this type of missingness as well.

In Sect. 34.7, two approaches to modeling data with non-random dropout are considered: the selection model framework (Sect. 34.7.1) and the pattern-mixture modeling family (Sect. 34.7.3). The optimal place of such missing not at random (MNAR) models is within a sensitivity analysis. One can use different plausible models within one of the modeling frameworks or across both. In Sect. 34.7.5, local influence is considered, which is a sensitivity analysis tool that can be used within the selection model framework.

The case study that was introduced in Sect. 34.2 is analyzed and discussed throughout the chapter, by way of running example.

34.2 The Age-Related Macular Degeneration Trial

In this section, the age-related macular degeneration trial is introduced. These data arise from a randomized multi-centric clinical trial comparing an experimental treatment (interferon-α) with a corresponding placebo in the treatment of patients with age-related macular degeneration. Here, we focus on the comparison between placebo and the highest dose (six million units daily) of interferon-α, but the full results of this trial have been reported elsewhere (Pharmacological Therapy for Macular Degeneration Study Group 1997). Patients with macular degeneration progressively lose vision. In the trial, the patients' visual acuity was assessed at different time points (4, 12, 24, and 52 weeks) through patients' ability to read lines of letters on standardized vision charts. These charts display lines of five letters of decreasing size, which the patient must read from top (largest letters) to bottom (smallest letters). The patient's visual acuity is the total number of letters correctly read. In addition, one often refers to each line with at least four letters correctly read as a "line of vision".

An endpoint of interest in this trial was the visual acuity at 1 year (treated as a continuous endpoint). Table 34.1 shows the visual acuity recorded as the number of letters read (mean and standard error) by treatment group at baseline, and at the four measurement occasions after baseline. Another point of interest is whether the visual acuity increases or decreases compared to baseline. Accordingly, this dichotomous response variable is created. Table 34.2 displays the percentages of patients with such a decrease of visual acuity compared to baseline at all four time points, both overall and by treatment arm.

The original quasi-continuous outcome, visual acuity, as well as the binary indicator for increase or decrease in number of letters read compared to baseline will be analyzed throughout this chapter.

Let us now graphically explore these data. The average evolution describes how the mean profile for a number of relevant subpopulations (or the population as a whole) evolves over time. The individual profiles are displayed in Fig. 34.1, while the mean profiles of the visual acuity values per treatment arm, as well as their 95% confidence intervals, are plotted in Fig. 34.2.

The average profiles indicate a decrease over time, which seems equally strong for both groups since the profiles are nearly parallel. However, the average visual

Table 34.1 Age-related macular degeneration trial. Mean (standard error) of visual acuity at baseline, at 4, 12, 24, and 52 weeks according to randomized treatment group (placebo versus interferon-α)

Time point	Placebo	Interferon-α	Total
Baseline	55.3 (1.4)	54.6 (1.4)	55.0 (1.0)
4 weeks	54.0 (1.5)	50.9 (1.5)	52.5 (1.1)
12 weeks	52.9 (1.6)	48.7 (1.7)	50.8 (1.2)
24 weeks	49.3 (1.8)	45.5 (1.8)	47.5 (1.3)
1 year (52 weeks)	44.4 (1.8)	39.1 (1.9)	42.0 (1.3)

Table 34.2 Age-related macular degeneration trial. Percentage of patients with a decrease of visual acuity compared to baseline, at 4, 12, 24, and 52 weeks according to randomized treatment group (placebo versus interferon-α)

Time point	Placebo	Interferon-α	Total
4 weeks	65.2%	72.1%	68.6%
12 weeks	59.6%	73.6%	66.4%
24 weeks	65.8%	74.3%	69.8%
1 year (52 weeks)	77.5%	83.7%	80.3%

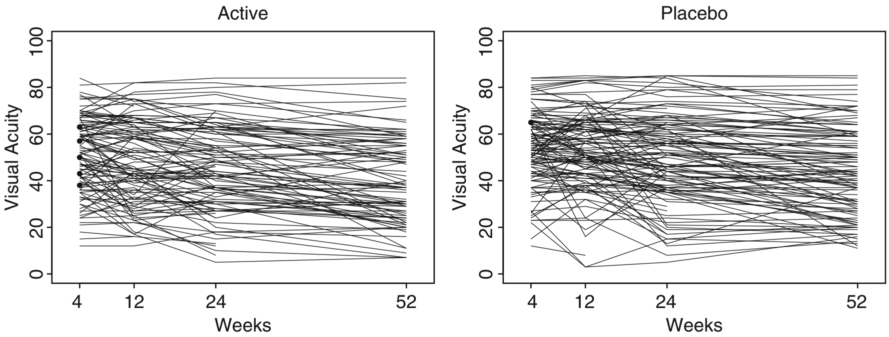


Fig. 34.1 Age-related macular degeneration trial. Individual profiles by treatment arm

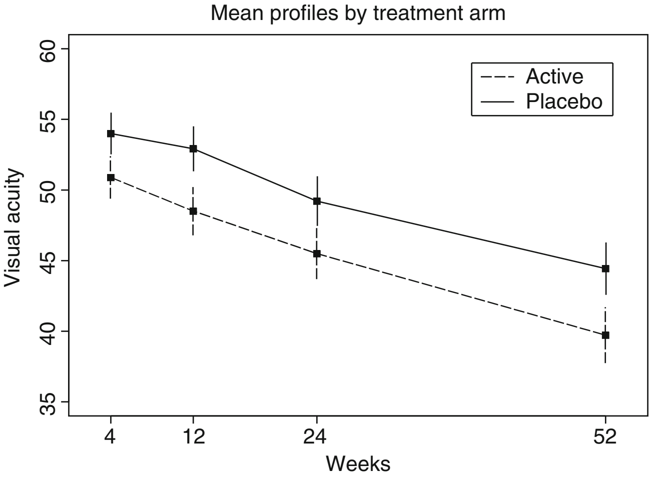


Fig. 34.2 Age-related macular degeneration trial. Mean profiles and 95% confidence intervals by treatment arm

acuity for the active group is overall smaller compared to the placebo group. As can be seen from the confidence intervals, these differences are clearly not significant at week 4 and 24, but might be for week 12 and 52. The individual profiles augment the averaged plot with a suggestion of the variability seen within the data.

Table 34.3 Age-related macular degeneration trial. Overview of missingness patterns and the frequencies with which they occur

Measurement occasion					
Week 4	Week 12	Week 24	Week 52	Number	%
Completers					
O	O	O	O	188	78.33
Dropouts					
O	O	O	M	24	10.00
O	O	M	M	8	3.33
O	M	M	M	6	2.50
M	M	M	M	6	2.50
Non-monotone missingness					
O	O	M	O	4	1.67
O	M	M	O	1	0.42
M	O	O	O	2	0.83
M	O	M	M	1	0.42

O observed, *M* missing

Thinning of the data towards the later study times suggests that trends at later times should be treated with caution. Indeed, an extra level of complexity is added whenever not all planned measurements are observed. This results in *incompleteness* or *missingness*. Another frequently encountered term is *dropout*, which refers to the case where all observations on a subject are obtained until a certain point in time, after which all measurements are missing. To simplify matters, we will largely focus on dropout, but a lot of the developments made are valid also for general types of missingness.

Regarding missingness in the ARMD data set, an overview of the different missingness patterns is given in Table 34.3. The missingness pattern shows at which of the scheduled time points the patient has an observation or a missing value. Clearly, both intermittent missingness as well as dropout occurs. It is observed that 188 of the 240 profiles are complete, which is a percentage of 78.33%, while 18.33% (44 subjects) drop out before the end of the study. Out of the latter group, 2.50% or six subjects have no follow-up measurements. The remaining 3.34%, representing eight subjects, have intermittent or non-monotone missing values. Although the group of dropouts is of considerable magnitude, the ones with intermittent missingness is much smaller. To be consistent throughout the analysis in this chapter, subjects with non-monotone profiles will not be taken into account. Further, neither will the subjects without any follow-up measurement be included in the analysis. Both reasons will become clear in the course of this chapter. This results in a data set of 226 of 240 subjects, or 94.17%.

One issue, resulting from dropout, is evidently a depletion of the study subjects. Of course, a decreasing sample size increases variability which, in turn, decreases precision. Note that the treatment might be a potential factor that could influence a patient's probability of dropping out. Although a large part of the scientist's study

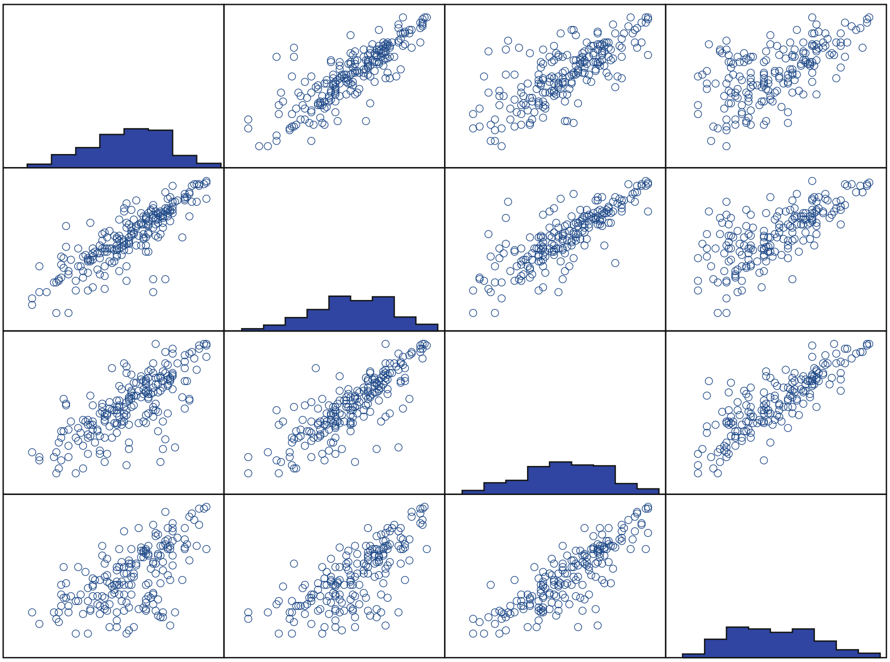


Fig. 34.3 Age-related macular degeneration trial. Scatter plot matrix for all four time points

interest will typically focus on the treatment effect, we should be aware that it is still a covariate and hence a design factor. Another question that will arise is whether dropout depends on observed or unobserved responses.

A different way of displaying several structural aspects is using a scatter plot matrix, such as in Fig. 34.3. The off-diagonal elements picture scatter plots of standardized residuals obtained from pairs of measurement occasions. The decay of correlation with time is studied by considering the evolution of the scatters with increasing distance to the main diagonal. Stationarity, on the other hand, implies that the scatter plots remain similar within diagonal bands *if measurement occasions are approximately equally spaced*. In addition to the scatter plots, we place histograms on the diagonal, capturing the variance structure. Features such as skewness, multimodality, and so forth, can then be graphically detected.

Further, the variance function is displayed in Fig. 34.4. The variance function is increasing non-linearly over time, and hence an unstructured variance model is a plausible starting point.

To gain further insight into the impact of dropout, it is useful to construct dropout-pattern-specific plots. Figure 34.5 displays the average profiles per dropout pattern, for both the interferon-α (active) and the placebo group.

This plot will be useful as a graphical start to a so-called pattern-mixture analysis, as described later in Sect. 34.7.3. The individual profiles plot, by definition

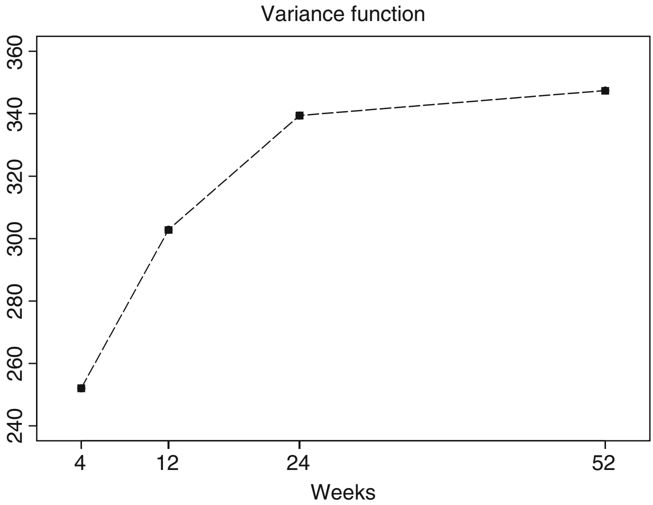


Fig. 34.4 Age-related macular degeneration trial. Variance function

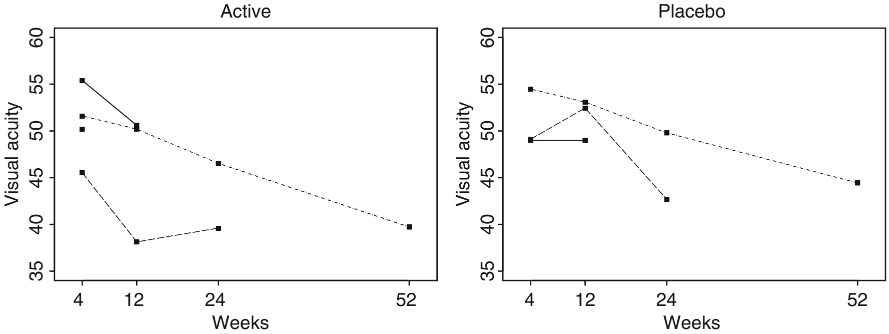


Fig. 34.5 Age-related macular degeneration trial. Mean profiles per pattern for both treatment arms

displaying all available data, has some intrinsic limitations and hence will not be considered. Indeed, as is the case with any individual data plot, it tends to be fairly busy. Further, if the same time axis is used for all profiles, also for those that drop out early, very little information can be extracted. In addition, the eye assigns more weight to the longer profiles, even though, they are considerably less frequent. In exploring the average profiles per pattern still care has to be taken for not overinterpreting the longer profiles and neglecting the shorter profiles.

Several observations can be made. In contrast to the average profile of the completers, the profiles of the incompleters do not show a linear trend (accept for the obvious case of two measurements). This implies that the impression from all patterns together may differ radically from a pattern-specific look. These conclusions seem to be consistent across treatment arms.

Another important observation is that those who drop out rather early, after week 12, increase from the start for the active group, and remain relatively stable for the placebo group. On the other hand patients who have the last measurement missing and are in the active group show a steep decrease between week 4 and 12, followed by a slight increase up to week 24. For the placebo group it is the other way around, the average profile increases a little from week 4–12, but decreases afterward. This implies that the average evolution is different for both treatment groups and also across the dropout patterns. Looked upon from the standpoint of dropout, this suggests that dropout will probably be related to the (observed) visual acuity values. Additionally, there is at least one important characteristic that makes dropout increase, namely, an unfavorable (downward) evolution of the visual acuity.

Arguably, careful modeling of these data, irrespective of the approach chosen, should reflect these features. We will consider the most important routes typically taken, starting from simple ad hoc methods, and then going to more principled methods, for which we first need to develop a formal but intuitively appealing framework.

34.3 Fundamental Concepts of Incomplete Multivariate Data

Data arising from epidemiological studies and biomedical research in general are often prone to incompleteness, or missingness. Since missingness usually occurs outside the control of the investigators, and may be related to the outcome of interest, it is generally necessary to address the process that governs this incompleteness. Only in special but important cases is it possible to ignore the missingness process.

In this section, we first introduce some general concepts of modeling incomplete data. Next, we discuss methods to model longitudinal or multivariate data both in the Gaussian and non-Gaussian setting, which capture the methodology necessary for the ARMD case study. For continuous measurements, the linear mixed model is used, whereas for non-Gaussian outcomes we distinguish between three model families: marginal, random-effects, and conditional models.

34.3.1 A Framework for Handling Missing Values

The nature of the missingness mechanism can affect the analysis of incomplete data and its resulting statistical inference. Therefore, we will introduce the terminology and notation necessary when modeling incomplete data, as well as the different missing data mechanisms.

Let the random vector $\vec{Y}$ correspond to the, possibly notional, complete set of measurements on a subject whether observed or not and suppose that its distribution depends on a vector of parameters $\vec{\theta}$. Let $\vec{R}$ be the associated missing value indicator, with distribution depending on parameter vector $\vec{\psi}$. In case missingness is restricted to dropout, we define D to be the occasion of dropout. For a particular realization of this pair $(\vec{y}, \vec{r})$, the elements of $\vec{r}$ take the values 1 and 0 indicating, respectively, whether the corresponding values of $\vec{y}$ are observed or not. When we are dealing with dropout the information in $\vec{r}$ can be summarized in a single variable: the first time at which a value is missing. Let $(\vec{y}_o, \vec{y}_m)$ denote the partition of $\vec{y}$ into the respective sets of observed and missing data. In what follows we will be attempting to fit and make inferences about models constructed for the pair $(\vec{y}, \vec{r})$ using only the observed data: $(\vec{y}_o, \vec{r})$. If $f(\vec{y}, \vec{r})$ is the joint distribution of the complete data, then the marginal distribution of the observed data, which forms the basis for model fitting, is

$$f(\vec{y}_o, \vec{r}) = \int f(\vec{y}, \vec{r})\dot{\vec{y}}_m. \tag{34.1}$$

Rubin's taxonomy (Rubin 1976; Little and Rubin 1987) of missing value processes is fundamental to modeling incomplete data. This classification essentially distinguishes settings in which important simplifications of this process are possible.

34.3.1.1 Missing Completely at Random (MCAR)

Under an MCAR mechanism the probability of an observation being missing is independent of the responses: $\mathrm{P}(\vec{R} = \vec{r} \mid \vec{y}) = \mathrm{P}(\vec{R} = \vec{r})$. The joint distribution of the *observed* data partitions as follows: $f(\vec{y}_o, \vec{r}) = f(\vec{y}_o; \vec{\theta})\mathrm{P}(\vec{r}; \vec{\psi})$. Under MCAR, the observed data can be analyzed as though the pattern of missing values was predetermined. In whatever way the data are analyzed, whether using a frequentist or likelihood procedure, the process generating the missing values can be ignored. For example, in this situation simple averages of the observed data at different times provide unbiased estimates of the underlying marginal profiles.

34.3.1.2 Missing at Random (MAR)

Under an MAR mechanism, the probability of an observation being missing is *conditionally* independent of the unobserved data, given the values of the observed data: $\mathrm{P}(\vec{R} = \vec{r} \mid \vec{y}) = \mathrm{P}(\vec{r} \mid \vec{y}_o)$, and again the joint distribution of the observed data can be partitioned:

$$f(\vec{y}_o, \vec{r}) = f(\vec{y}_o; \vec{\theta})\mathrm{P}(\vec{r} \mid \vec{y}_o; \vec{\psi}). \tag{34.2}$$

An example of random dropout occurs in a trial in which subjects are removed when their observed response drifts outside prescribed limits, Murray and Findlay (1988) describe an instance of this. We note that the handling of the MAR assumption is much easier with dropout than with more general missing value patterns.

No mention has been made of covariates and the dependence of missing value probabilities on these. According to Little (1995), the original intention was that MCAR should refer to the case in which the probability of a value being missing does not depend on the response *or* covariates, and suggests the term "covariate-dependent" missing value mechanism for the case where it depends on the latter.

This avoids the problem of having the class of mechanisms potentially changing with the addition and removal of covariates.

One can see the importance of the MAR assumption from an intuitive viewpoint. Essentially it states that once appropriate account is taken of what we have observed, there remains no dependence on unobservables, at least in terms of the probability model. We should as a consequence expect much of the missing value problem to disappear under the MAR mechanism and this is in fact the case. This can be shown more formally through consideration of the likelihood. The result (Eq. 34.2) implies that the joint log-likelihood for $\vec{\theta}$ and $\vec{\psi}$ partitions:

$$\ell(\vec{\theta}, \vec{\psi}; \vec{y}_o, \vec{r}) = \ell(\vec{\theta}; \vec{y}_o) + \ell(\vec{\psi}; \vec{r}).$$

Provided that $\vec{\theta}$ and $\vec{\psi}$ are not interdependent, information about the response model parameter $\vec{\theta}$ is contained wholly in $\ell(\vec{\theta}; \vec{y}_o)$, the log-likelihood of the observed response data. This is the log-likelihood function that is used when no account is taken of the missing value mechanism, hence for a likelihood analysis under the MAR assumption, the missing value mechanism is said to be *ignorable*. It should be noted that although the correct maximum likelihood estimates and likelihood ratio statistics will be generated by the use of $\ell(\vec{\theta}; \vec{y}_o)$, some care needs to be taken with the choice of appropriate sampling distribution in a frequentist analysis. For this the missing value mechanism is *not* ignorable, even under MAR (Kenward and Molenberghs 1998). In practice though there is little reason for worry since this just means that estimates of precision should be based on the observed rather than the expected information matrix. Additionally, it also has been shown how non-likelihood approaches can be developed for the MAR case (Robins et al. 1995, 1998; Fitzmaurice et al. 1995).

While the MAR assumption is particularly convenient in that it leads to considerable simplification in the issues surrounding the analysis of incomplete longitudinal data, it is rare in practice to be able to justify its adoption, and so in many situations the final class of missing value mechanisms applies.

34.3.1.3 Missing Not at Random (MNAR)

In this case neither MCAR nor MAR hold. Under MNAR, the probability of a measurement being missing depends on unobserved data. No simplification of the joint distribution is possible and inferences can only be made by making further assumptions, about which the observed data alone carry no information. Ideally the choice of such assumptions should be guided by external information, but the degree to which this is possible in practice varies greatly. In attempting to formulate models for the joint distribution of the response and dropout process, $f(\vec{y}, \vec{r})$, three main types of model have been used and these are defined by two possible factorizations of this distribution. The first, the *selection* model, is based on

$$f(\vec{y}, \vec{r}) = f(\vec{y})\mathrm{P}(\vec{r} \mid \vec{y}). \tag{34.3}$$

The second, the *pattern-mixture* model (PMM), uses

$$f(\vec{y}, \vec{r}) = f(\vec{y} \mid \vec{r})\mathrm{P}(\vec{r}). \tag{34.4}$$

The differences between these reversed factorizations are important in the MNAR case, and lead to quite different, but complementary, views of the missing value problem. Little (1995) and Hogan and Laird (1997) provide detailed reviews. The term "selection model" originates from the econometric literature (Heckman 1976) and it can be seen that a subject's missing values are "selected" through the probability model, given their measurements, whether observed or not. Rubin's classification is defined in the selection framework, and imposition of conditions on $\mathrm{P}(\vec{r} \mid \vec{y})$ determines to which of the three classes the model belongs in the frequentist sense. On the other hand, the pattern-mixture model allows a different response model for each pattern of missing values, the observed data being a mixture of these weighted by the probability of each missing value or dropout pattern. At first sight such a model is less appealing in terms of probability mechanisms for generating the data, but it has other important advantages. Recently it has been shown, for dropout, how the Rubin classification can be applied in the pattern-mixture framework as well (Molenberghs et al. 1998; Kenward et al. 2003).

Instead of using the selection or pattern-mixture model frameworks, the measurement and dropout process can be modeled jointly using a *shared-parameter* model (Wu and Carroll 1988; Wu and Bailey 1988, 1989; TenHave et al. 1998). In such a model the measurement and dropout process are assumed to be independent, conditional upon a certain set of shared parameters. This shared-parameter model is formulated by way of the following factorization:

$$f(\vec{y}, \vec{r}) = f(\vec{y} \mid \vec{b}_i)\mathrm{P}(\vec{r} \mid \vec{b}_i). \tag{34.5}$$

Here, $\vec{b}_i$ are shared parameters, often considered to be random effects and following a specific parametric distribution.

We will consider the first two approaches to modeling data with non-random dropout in Sects. 34.7.1 and 34.7.3.

These strategies used to analyze incomplete data are based on two choices. First, a choice has to be made regarding the modeling approach to the measurement sequence. The measurement model will depend on whether or not a full longitudinal analysis is done. In case focus is on the last observed measurement or on the last measurement occasion only, one typically opts for classical two- or multi-group comparisons (t test, Wilcoxon, etc.). In case a longitudinal analysis is deemed necessary, the choice made depends on the nature of the outcome. A variety of model both for Gaussian and non-Gaussian multivariate data will be discussed in Sect. 34.3.2.

Second, a choice has to be made regarding the modeling approach for the missingness process. Luckily, under certain assumptions this process can be ignored. Indeed, Rubin (1976) has shown that, under MAR and mild regularity conditions (parameters $\vec{\theta}$ and $\vec{\psi}$ are functionally independent), likelihood-based inference is valid when the missing data mechanism is ignored (see also Verbeke and Molenberghs 2000; Molenberghs and Kenward 2007). Practically speaking, the likelihood of interest is then based upon the factor $f(\vec{y}_i^o|\vec{\theta})$. This is called *ignorability*. The practical implication is that a software module with likelihood estimation facilities and with the ability to handle incompletely observed subjects manipulates the correct likelihood, providing valid parameter estimates and likelihood ratio values.

A few cautionary remarks are in place. First, when at least part of the scientific interest is directed toward the non-response process, obviously both processes need to be considered. Still, under MAR, both processes can be modeled and parameters estimated separately. Second, likelihood inference is often surrounded with references to the sampling distribution (e.g., to construct precision estimators and for statistical hypothesis tests; Kenward and Molenberghs 1998). However, the practical implication is that standard errors and associated tests, when based on the observed rather than the expected information matrix and given the parametric assumptions are correct, are valid. Third, it may be hard to fully rule out the operation of an MNAR mechanism. This point was brought up in the introduction and will be discussed further in Sect. 34.7. Fourth, an analysis can proceed only when a full longitudinal analysis is necessary, even when interest lies, for example, in a comparison between the two treatment groups at the last occasion. In the latter case, the fitted model can be used as the basis for inference at the last occasion.

Assume that $\vec{Y}_i$ is the n-dimensional response vector for subject i, containing the outcomes at n various measurement occasions, and $1 \leq i \leq N$, with N is the number of subjects. Note that in general, the number of measurements occasions can differ among patients, which would yield n_i instead of n. Further assume that incompleteness is due to dropout only, and that the first measurement Y_{i1} is obtained for everyone. Under the selection model framework, a possible model for the dropout process is based on a logistic regression for the probability of dropout at occasion j, given the subject is still in the study. We denote this probability by $g(\vec{h}_{ij}, y_{ij})$ in which $\vec{h}_{ij}$ is a vector containing all responses observed up to but not including occasion j, as well as relevant covariates. We then assume that $g(\vec{h}_{ij}, y_{ij})$ satisfies

$$\text{logit}\left[g(\vec{h}_{ij}, y_{ij})\right] = \text{logit}\left[\text{pr}\left(D_i = j|D_i \geq j, \vec{y}_i\right)\right] = \vec{h}_{ij}\vec{\psi} + \omega y_{ij}. \quad (34.6)$$

When ω equals zero, the dropout model is MAR, and all parameters can be estimated using standard software since the measurement model and the dropout model can then be fitted separately. If $\omega \neq 0$, the posited dropout process is MNAR. Model (34.6) provides the building blocks for the dropout process $f(d_i|\vec{y}_i, \vec{\psi})$. This model is often referred to as Diggle and Kenward's (1994) model. A review of this Diggle–Kenward model is provided in Sect. 34.7.1.

34.3.2 A Framework for Longitudinal or Multivariate Data

Let us now turn attention to standard model frameworks for longitudinal data, or multivariate data in general, which will be used to analyze the ARMD case study. First, the continuous case will be treated where the linear mixed model undoubtedly occupies the most prominent role. Then, we switch to the discrete setting, where important distinctions exist between three model families: the marginal, random-effects, and conditional model family. The mixed model parameters, both in the continuous and discrete cases, are usually estimated using maximum-likelihood-based methods, which implies that the results are valid under the MAR assumption. In contrast, generalized estimating equations, which is a commonly encountered marginal approach to non-Gaussian data, have a frequentist foundation. Consequently, this method is valid only under the MCAR assumption, necessitating the need for extensions. This will be discussed in Sects. 34.5.2 and 34.5.3. Note that even though Bayesian methodology is important, the focus in this is on likelihood- and frequentist-based methodology. An overview of Bayesian methodology can be found in Ibrahim and Molenberghs (2009).

Laird and Ware (1982) proposed, for continuous outcomes, likelihood-based mixed-effects models. A broad discussion of such models is provided in Verbeke and Molenberghs (2000). The general linear mixed effects model, perhaps with serial correlation, is the following:

$$\vec{Y}_i = X_i \vec{\beta} + Z_i \vec{b}_i + \vec{W}_i + \vec{\varepsilon}_i, \tag{34.7}$$

where $\vec{Y}_i$ is the n-dimensional response vector for subject i, $1 \le i \le N$, N is the number of subjects, X_i and Z_i are $(n \times p)$ and $(n \times q)$ known design matrices, $\vec{\beta}$ is the p-dimensional vector containing the fixed effects, $\vec{b}_i \sim N(\vec{0}, D)$ is the q-dimensional vector containing the random effects, $\vec{\varepsilon}_i \sim N(\vec{0}, \sigma^2 I_n)$ is a n-dimensional vector of measurement error components, and $\vec{b}_1, \ldots, \vec{b}_N, \vec{\varepsilon}_1, \ldots, \vec{\varepsilon}_N$ are assumed to be independent. Serial correlation is captured by the realization of a Gaussian stochastic process, $\vec{W}_i$, which is assumed to follow a $N(\vec{0}, \tau^2 H_i)$ law. The serial covariance matrix H_i only depends on i through the time points t_{ij} at which measurements are taken. The structure of the matrix H_i is determined through the autocorrelation function $\rho(t_{ij} - t_{ik})$. This function decreases such that $\rho(0) = 1$ and $\rho(+\infty) = 0$. Finally, D is a general $(q \times q)$ covariance matrix with (i, j) element $d_{ij} = d_{ji}$. Inference is based on the marginal distribution of the response $\vec{Y}_i$ which, after integrating over random effects, can be expressed as

$$\vec{Y}_i \sim N\left(X_i \vec{\beta}, Z_i D Z_i' + \Sigma_i\right). \tag{34.8}$$

Here, $\Sigma_i = \sigma^2 I_n + \tau^2 H_i$ is a $(n \times n)$ covariance matrix grouping the measurement error and serial components.

Two popular choices to capture serial correlation is by means of exponential or Gaussian decay. An exponential process is based on writing the correlation between two residuals at times t_{ij} and t_{ik} as

$$\mathrm{Corr}\left(t_{ij}, t_{ik}\right) = \exp\left(-\frac{|t_{ij} - t_{ik}|}{\phi}\right) = \rho^{|t_{ij}-t_{ik}|}, \tag{34.9}$$

($\phi > 0$), where $\rho = \exp(-1/\phi)$. The Gaussian counterpart is

$$\mathrm{Corr}\left(t_{ij}, t_{ik}\right) = \exp\left(-\frac{\left(t_{ij} - t_{ik}\right)^2}{\phi^2}\right) = \rho^{\left(t_{ij}-t_{ik}\right)^2}, \tag{34.10}$$

($\phi > 0$), $\rho = \exp(-1/\phi^2)$. It follows from Eq. 34.7 that, conditional on the random effect $\vec{b}_i$, $\vec{Y}_i$ is normally distributed with mean vector $X_i\vec{\beta} + Z_i\vec{b}_i$ and with covariance matrix Σ_i. Define $V_i = Z_i D Z_i' + \Sigma_i$, then the marginal distribution of $\vec{Y}_i$ is $\vec{Y}_i \sim N(X_i\vec{\beta}, V_i)$.

In a clinical-trial setting and in some epidemiological settings, one often has balanced data in the sense that the measurement occasions are common to all patients, as we assume here. In such a case, one often considers the random effects as nuisance parameters. The focus is then, for example, on the marginal compound-symmetry model rather than on the hierarchical random-intercepts model. Or, more generally, one considers $\Sigma_i = \Sigma$ to be unstructured and then no random effects are explicitly included. Especially when the number of patients is much larger than the number of measurement occasions within a patient, such an approach is useful. In an epidemiological setting however, one is often confronted with longer measurement sequences, perhaps also unequally spaced and/or of unequal length.

Whereas the linear mixed model is seen as a unifying parametric framework for Gaussian repeated measures (Verbeke and Molenberghs 2000), there are a variety of methods in common use in the non-Gaussian setting. Several authors, such as Diggle et al. (2002) and Aerts et al. (2002) distinguish between three such families. Still focusing on continuous outcomes, a marginal model is characterized by the specification of a marginal mean function

$$E\left(Y_{ij}|\vec{x}_{ij}\right) = \vec{x}_{ij}'\vec{\beta}, \tag{34.11}$$

whereas in a random-effects model we focus on the expectation, conditional upon the random-effects vector:

$$E\left(Y_{ij}|\vec{b}_i, \vec{x}_{ij}\right) = \vec{x}_{ij}'\vec{\beta} + \vec{z}_{ij}'\vec{b}_i. \tag{34.12}$$

Finally, a third family of models conditions a particular outcome on the other responses or a subset thereof. In particular, a simple first-order stationary transition model focuses on expectations of the form

$$E\left(Y_{ij}|Y_{i,j-1}, \ldots, Y_{i1}, \vec{x}_{ij}\right) = \vec{x}_{ij}'\vec{\beta} + \alpha Y_{i,j-1}. \tag{34.13}$$

In the linear mixed model case, random-effects models imply a simple marginal model. This is due to the elegant properties of the multivariate normal distribution. In particular, the expectation Eq. 34.11 follows from Eq. 34.12 by either (1) marginalizing over the random effects or by (2) conditioning upon the random-effects vector $\vec{b}_i = \vec{0}$. Hence, the fixed-effects parameters $\vec{\beta}$ have both a marginal as well as a hierarchical model interpretation. Finally, when a conditional model is expressed in terms of residuals rather than outcomes directly, it also leads to particular forms of the general linear mixed effects model.

Such a close connection between the model families does not exist when outcomes are of a non-normal type, such as binary, categorical, or discrete. We will consider each of the model families in turn and then point to some particular issues arising within them or when comparisons are made between them. We focus on the first two important subfamilies of models in turn, the marginal and random-effects models.

34.3.3 Marginal Models

In marginal models, the parameters characterize the marginal probabilities of a subset of the outcomes, without conditioning on the other outcomes. Thorough discussions on marginal modeling can be found in Diggle et al. (2002) and in Molenberghs and Verbeke (2005). Even though a variety of flexible full-likelihood models exist, maximum likelihood can be unattractive due to excessive computational requirements, especially when high-dimensional vectors of correlated data arise. As a consequence, alternative methods have been in demand. One such non-likelihood approach which is commonly used is *generalized estimating equations* (Liang and Zeger 1986) and will be discussed below.

Generalized Estimating Equations For clustered and repeated data, Liang and Zeger (1986) proposed so-called *generalized estimating equations* (GEE) which require only the correct specification of the univariate marginal distributions provided one is willing to adopt "working" assumptions about the association structure. They estimate the parameters associated with the expected value of an individual's vector of binary responses and phrase the working assumptions about the association between pairs of outcomes in terms of marginal correlations. The method combines estimating equations for the regression parameters $\vec{\beta}$ with moment-based estimating for the correlation parameters entering the working assumptions.

Let us now explain the GEE methodology a little further. The score equations to be solved when computing maximum likelihood estimates under a marginal generalized linear model (34.11) for non-Gaussian correlated outcomes are given by

$$\sum_{i=1}^{N} \frac{\partial \vec{\mu}_i}{\partial \vec{\beta}'} V_i^{-1} \left(\vec{y}_i - \vec{\mu}_i \right) = \vec{0}, \tag{34.14}$$

where $\vec{\mu}_i = E(\vec{y}_i)$ and V_i is the so-called working covariance matrix, that is, V_i approximates $\text{Var}(\vec{Y}_i)$, the true underlying covariance matrix for $\vec{Y}_i$. This working covariance matrix can be decomposed as $V_i = A_i^{1/2} R_i A_i^{1/2}$, in which $A_i^{1/2}$ is again the diagonal matrix with the marginal standard deviations of $\vec{Y}_i$ along the diagonal, and $R_i = \text{Corr}(\vec{Y}_i)$ is the correlation matrix. Usually, the marginal correlation matrix R_i contains a vector $\vec{\alpha}$, $R_i = R_i(\vec{\alpha})$ of unknown parameters which is replaced for practical purposes by a consistent estimate. Liang and Zeger (1986) dealt with this set of nuisance parameters $\vec{\alpha}$ by allowing for specification of an incorrect structure for R_i or so-called working correlation matrix. Assuming that the marginal mean $\vec{\mu}_i$ has been correctly specified as $h(\vec{\mu}_i) = X_i \vec{\beta}$, they further showed that, under mild regularity conditions, the estimator $\widehat{\vec{\beta}}$ obtained from solving Eq. 34.14 is asymptotically normally distributed with mean $\vec{\beta}$ and with covariance matrix

$$I_0^{-1} I_1 I_0^{-1}, \tag{34.15}$$

where

$$I_0 = \left(\sum_{i=1}^{N} \frac{\partial \vec{\mu}_i'}{\partial \vec{\beta}} V_i^{-1} \frac{\partial \vec{\mu}_i}{\partial \vec{\beta}'} \right), \tag{34.16}$$

$$I_1 = \left(\sum_{i=1}^{N} \frac{\partial \vec{\mu}_i'}{\partial \vec{\beta}} V_i^{-1} \text{Var}\left(\vec{y}_i\right) V_i^{-1} \frac{\partial \vec{\mu}_i}{\partial \vec{\beta}'} \right). \tag{34.17}$$

Consistent precision estimates can be obtained by replacing all unknown quantities in Eq. 34.15 by consistent parameter estimates. Observe that, when R_i is correctly specified, $\text{Var}(\vec{Y}_i) = V_i$ in Eq. 34.17, and thus $I_0 = I_1$. As a result, the expression for the covariance matrix (Eq. 34.15) reduces to I_0^{-1}, corresponding to full likelihood, that is, when the first and second moment assumptions are correct. Thus, when the working correlation structure is correctly specified, it reduces to full likelihood, although generally it differs from it.

On the other hand, when the working correlation structure differs strongly from the true underlying structure, there is no price to pay in terms of consistency of the asymptotic normality of $\widehat{\vec{\beta}}$, but such a poor choice may result in loss of efficiency. With incomplete data that arise under the MAR or MNAR assumption, an erroneously specified working correlation matrix may additionally lead to bias (Molenberghs and Kenward 2007).

Two further specifications are necessary before GEE is operational: $\text{Var}(\vec{Y}_i)$ on the one hand and $R_i(\vec{\alpha})$, with estimation of $\vec{\alpha}$ in particular, on the other hand. In practice, $\text{Var}(\vec{y}_i)$ in (34.15) is replaced by $(\vec{y}_i - \vec{\mu}_i)(\vec{y}_i - \vec{\mu}_i)'$, which is unbiased on the sole condition that the mean was again correctly specified. Second, one also needs estimates of the nuisance parameters of $\vec{\alpha}$. Liang and Zeger (1986) proposed moment-based estimates for the working correlation.

34.3.4 Random-Effects Models

Models with random effects are differentiated from marginal models by the inclusion of parameters which are specific to the cluster or subject. Unlike for correlated Gaussian outcomes, the parameters of the random effects and marginal models for correlated non-Gaussian data describe different types of effects of the covariates on the response probabilities (Neuhaus 1992). The choice between marginal and random-effects strategies should heavily depend on the scientific goals. Marginal models evaluate the overall risk as a function of covariates. With a random-effects approach, the response rates are modeled as a function of covariates and parameters, specific to a cluster or subject. In such models, interpretation of fixed-effect parameters is conditional on a constant level of the random-effects parameter. Marginal comparisons, on the other hand, make no use of within-cluster comparisons for cluster varying covariates and are therefore not useful to assess within-subject effects (Neuhaus et al. 1991).

While several non-equivalent random-effects models exist, one of the most popular is the *generalized linear mixed model* (Breslow and Clayton 1993) which is closely related to linear and non-linear mixed models, which is discussed in the following subsection.

Generalized Linear Mixed Models A general framework for mixed-effects models can be expressed as follows. Assume that $\vec{Y}_i$ (possibly appropriately transformed) satisfies

$$\vec{Y}_i | \vec{b}_i \sim F_i(\vec{\theta}, \vec{b}_i), \tag{34.18}$$

that is, conditional on $\vec{b}_i$, $\vec{Y}_i$ follows a prespecified distribution F_i, possibly depending on covariates, and parameterized through a vector $\vec{\theta}$ of unknown parameters, common to all subjects. Further, $\vec{b}_i$ is a q-dimensional vector of subject-specific parameters, called random effects, assumed to follow a so-called mixing distribution G which may depend on a vector $\vec{\psi}$ of unknown parameters, that is, $\vec{b}_i \sim G(\vec{\psi})$. The $\vec{b}_i$ reflect the between-unit heterogeneity in the population with respect to the distribution of $\vec{Y}_i$. In the presence of random effects, conditional independence is often assumed, under which the components Y_{ij} in $\vec{Y}_i$ are independent, conditional on $\vec{b}_i$. The distribution function F_i in Eq. 34.18 then becomes a product over the n independent elements in $\vec{Y}_i$.

In general, unless a fully Bayesian approach is followed, inference is based on the marginal model for $\vec{Y}_i$ which is obtained from integrating out the random effects, over their distribution $G(\vec{\psi})$. Let $f_i(\vec{y}_i|\vec{b}_i)$ and $g(\vec{b}_i)$ denote the density functions corresponding to the distributions F_i and G, respectively, then the marginal density function of $\vec{Y}_i$ equals

$$f_i(\vec{y}_i) = \int f_i\left(\vec{y}_i|\vec{b}_i\right) g(\vec{b}_i)\vec{b}_i, \tag{34.19}$$

which depends on the unknown parameters $\vec{\theta}$ and $\vec{\psi}$. Assuming independence of the units, estimates of $\widehat{\vec{\theta}}$ and $\widehat{\vec{\psi}}$ can be obtained from maximizing the likelihood function built from Eq. 34.19, and inferences immediately follow from classical maximum likelihood theory.

It is important to realize that the random-effects distribution G is crucial in the calculation of the marginal model (34.19). One often assumes G to be of a specific parametric form, such as a (multivariate) normal. Depending on F_i and G, the integration in Eq. 34.19 may or may not be possible analytically. Proposed solutions are based on Taylor series expansions of $f_i(\vec{y}_i|\vec{b}_i)$, or on numerical approximations of the integral, such as (adaptive) Gaussian quadrature. A general formulation of GLMM is as follows. Conditionally on random effects $\vec{b}_i$, it assumes that the elements Y_{ij} of $\vec{Y}_i$ are independent, with density function of the form

$$f\,(y|\theta_i,\phi) = \exp\left\{\phi^{-1}\left[y\theta_i - \psi(\theta_i)\right] + c(y,\phi)\right\},$$

with mean $E(Y_{ij}|\vec{b}_i)$ and variance $\mathrm{Var}(Y_{ij}|\vec{b}_i)$, and where, apart from a link function h, a linear regression model with parameters $\vec{\beta}$ and $\vec{b}_i$ is used for the mean, that is, $h(E(\vec{Y}_i|\vec{b}_i)) = X_i\vec{\beta} + Z_i\vec{b}_i$. Note that the linear mixed model is a special case, with identity link function. The random effects $\vec{b}_i$ are again assumed to be sampled from a (multivariate) normal distribution with mean $\vec{0}$ and covariance matrix D. When the link function is chosen to be of the logit form and the random effects are assumed to be normally distributed, the familiar logistic-linear GLMM follows.

There are at least two major differences in comparison to the linear mixed model discussed in the previous section. First, the marginal distribution of $\vec{Y}_i$ can no longer be calculated analytically, such that numerical approximations to the marginal density come into play, seriously complicating the computation of the maximum likelihood estimates of the parameters in the marginal model, i.e., $\vec{\beta}$, D, and the parameters in all Σ_i. A consequence is that the marginal covariance structure does not immediately follow from the model formulation, that is, it is not always clear in practice what assumptions a specific model implies with respect to the underlying variance function and the underlying correlation structure in the data.

A second important difference is with respect to the interpretation of the fixed effects $\vec{\beta}$. Under the linear model (34.7), $E(\vec{Y}_i)$ equals $X_i\vec{\beta}$, such that the fixed effects have a subject-specific as well as a population-averaged interpretation. Indeed, the elements in $\vec{\beta}$ reflect the effect of specific covariates, conditionally on the random effects $\vec{b}_i$, as well as marginalized over these random effects. Under non-linear mixed models, however, this does no longer hold in general. The fixed effects now only reflect the conditional effect of covariates, and the marginal effect is not easily obtained anymore as $E(\vec{Y}_i)$ which is given by

$$E(\vec{Y}_i) = \int \vec{y}_i \int f_i(\vec{y}_i|\vec{b}_i)g(\vec{b}_i)\mathrm{d}\vec{b}_i\mathrm{d}\vec{y}_i,$$

which, in general, is *not* of the form $h(E(Y_{ij}|\vec{b}_i = \vec{0})) = X_i\vec{\beta}$, that is, the link function of the conditional mean function evaluated in the zero random effects vector.

The non-linear nature of the model again implies that the marginal distribution of $\vec{y}_i$ is, in general, not easily obtained, such that model fitting requires approximation of the marginal density function. An exception to this occurs when the probit link is used. Further, as was also the case for non-linear mixed models, the parameters $\vec{\beta}$ have no marginal interpretation, except for some very particular models.

As an important example, consider the binomial model for binary data, with the logit canonical link function, and where the only random effects are intercepts b_i. It can then be shown that the marginal mean $\vec{\mu}_i = E(Y_{ij})$ satisfies $h(\vec{\mu}_i) \approx X_i\vec{\beta}^*$ with

$$\vec{\beta}^* = \left[c^2\text{Var}(b_i) + 1\right]^{-1/2}\vec{\beta}, \tag{34.20}$$

in which c equals $16\sqrt{3}/15\pi$. Hence, although the parameters $\vec{\beta}$ in the generalized linear mixed model have no marginal interpretation, they do show a strong relation to their marginal counterparts. Note that, as a consequence of this relation, larger covariate effects are obtained under the random-effects model in comparison to the marginal model.

34.4 Simple Ad Hoc Methods

Relatively simple methods that have been and still are in extensive use are valid when the measurement and missing data processes are independent and their parameters are separated, which is called the missing completely at random or MCAR assumption. Moreover, for some of these methods this assumption is necessary but not sufficient. It is important to realize that many of these methods are used also in situations where the MCAR assumption is not tenable. This should be seen as bad practice since it will often lead to biased estimates and invalid tests and hence to erroneous conclusions. Ample detail and illustrations of several problems are provided in Verbeke and Molenberghs (1997, Chap. 5).

Let us focus on two commonly used methods, complete case analysis and last observation carried forward (LOCF) analysis. The first one requires the strict MCAR assumption and is based on removing data, whereas LOCF requires even stronger and more unrealistic conditions to apply (Kenward and Molenberghs 2009). The LOCF method is discussed in the realm of single imputation techniques. For such single imputation methods, a single value is substituted for every "hole" in the data set and the resulting data set is analyzed as if it represented the true complete data. On the other hand, multiple imputation properly accounts for the uncertainty of imputation by multiply imputing the missing values given the observed ones (Rubin 1987; Schafer 1997). Additionally, it is valid under the less strict MAR assumption. The multiple imputation technique will be discussed in Sect. 34.5.3.

An often quoted advantage of the methods described below and related ones is that complete data software can be used. With the availability of such software like the SAS procedures GENMOD, MIXED, GLIMMIX, and NLMIXED, it is however no longer necessary to restrict oneself to complete data software, since they allow a likelihood-based ignorable analysis, using the data as they are, without deletion or imputation.

34.4.1 Complete Case Analysis

A complete case (CC) analysis includes only those cases for analysis, for which all measurements were recorded. This method has obvious advantages. It is very simple to describe and since the data structure is as would have resulted from a complete experiment, standard statistical software can be used. Further, since the entire estimation is done on the same subset of completers, there is a common basis for inference, unlike for the available case methods. Unfortunately, the method suffers from severe drawbacks. First, there is nearly always a substantial loss of information. For example, suppose there are 20 measurements, with 10% of missing data on each measurement. Suppose, further, that missingness on the different measurements is independent; then, the estimated percentage of incomplete observations is as high as 87%. The impact on precision and power is dramatic. Even though the reduction of the number of complete cases will be less dramatic in realistic settings where the missingness indicators are correlated, the effect just sketched will often undermine a complete case analysis. In addition, severe bias can result when the missingness mechanism is not MCAR. Indeed, should an estimator be consistent in the complete data problem, then the derived complete case analysis is consistent only if the missingness process is MCAR.

A simple partial check on the MCAR assumption is as follows (Little and Rubin 1987). Divide the observations on measurement j into two groups: (1) those subjects that are also observed on another measurement or set of measurements and (2) those missing on the other measurement(s). Should MCAR hold, then both groups should be random samples of the same population. Failure to reject equality of the distributional parameters of both samples increases the evidence for MCAR, but does not prove it.

34.4.2 Simple Forms of Imputation

An alternative way to obtain a data set on which complete data methods can be used is based on filling in rather than deletion. Commonly, the observed values are used to impute values for the missing observations. There are several ways to use the observed information. First, one can use information on the same subject (e.g., last observation carried forward). Second, information can be borrowed from other subjects (e.g., mean imputation). Finally, both within and between subject information can be used (e.g., conditional mean imputation, hot deck imputation). A standard reference is Little and Rubin (1987).

We will discuss the last observation carried forward method, which is the most commonly used single imputation technique.

Last Observation Carried Forward A method that has received a lot of attention (Siddiqui and Ali 1998; Mallinckrodt et al. 2003a,b) is last observation carried forward (LOCF). In the LOCF method, whenever a value is missing, the last observed value is substituted. The technique can be applied to both monotone and non-monotone missing data. It is typically applied to settings where incompleteness is due to attrition.

Very strong and often unrealistic assumptions have to be made to ensure validity of this method. First, one has to believe that a subjects' measurement stays at the same level from the moment of dropout onward (or during the period they are unobserved in the case of intermittent missingness). In a clinical trial setting, one might believe that the response profile *changes* as soon as a patient goes off treatment and even that it would flatten. However, the constant profile assumption is even stronger. Further, this method shares with other single imputation methods that it overestimates the precision by treating imputed and actually observed values on equal footing.

However, LOCF does not need to be seen as an imputation strategy. The situation in which the scientific question is in terms of the last observed measurement is often considered to be the real motivation for LOCF, though in some cases, the question defined as such may be perceived as having an unrealistic and ad hoc flavor. Clearly, measurements at (self-selected) dropout times are lumped together with measurements made at the (investigator defined) end of the study.

34.4.3 Discussion of Single Imputation Techniques

For review of other single imputation techniques we refer to Little and Rubin (1987) and Rubin (1987). They indicate great care has to be taken with these single imputations strategies. Dempster and Rubin (1983) write: "The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases." Indeed, all of these single imputation methods share the following pitfalls. First, since the imputation model could be wrong, the point estimates and comparisons of interest could be biased. Moreover, perhaps contrary to some common belief, such biases can be conservative (estimated effect is smaller in absolute value than the true one), but also liberal. Second, even for a correct imputation model, the uncertainty resulting from missingness is masked. Indeed, even when one is reasonably sure about the mean value the unknown observation *would have had*, the actual stochastic realization, depending on both the mean and error structures, is still unknown. Thus, the performance of imputation techniques is unreliable. In addition, these methods artificially increase the amount of information in the data by treating imputed and actually observed values on equal footing. This can lead to an artificially inflated precision. Finally, most methods require the MCAR assumption to hold while some even require additional and often unrealistically strong assumptions. Indeed, last observation carried forward requires the response to be unchanged until the end of the study.

34.5 Incomplete Data and MAR

As already pointed in Sect. 34.3.1, valid inference can be obtained under the MAR assumption through a likelihood-based analysis, without the need for modeling the dropout process. As a consequence, one can simply use, for example, linear or generalized linear mixed models as introduced in Sect. 34.3.2, without additional complication or effort. In this section, we will show there is no reason to use ad hoc methods when analysis valid under the MAR assumption can be implemented with standard software. Note that methods valid under MAR are also valid if data are MCAR, while the reverse does not hold.

34.5.1 Direct Likelihood

For longitudinal or multivariate data with missingness, a likelihood-based mixed-effects model only requires that the missing data are MAR. These mixed-effects models permit the inclusion of subjects with missing values at some time points (both dropout and intermittent missingness). For the continuous-outcome setting, this amounts to the general linear mixed model, introduced in Sect. 34.3.2, which can be viewed both in the marginal and hierarchical framework.

However, for the non-Gaussian case, such mixed-effects models are restricted to the random-effects model family. Here, we focus on the generalized linear mixed models, as also discussed in Sect. 34.3.2.

Such a likelihood-based MAR analysis is also termed likelihood-based ignorable analysis or direct-likelihood analysis. Obviously in such an analysis the observed data are used without deletion or imputation. In doing so, appropriate adjustments valid under MAR are made to parameters at times when data are incomplete, due to the within-patient correlation.

Settings in which the repeated measures are balanced, in the sense that a common set of measurement times is considered for all subjects, allow the a priori specification of a saturated model (e.g., full group by time interaction model for the fixed effects and unstructured variance-covariance matrix). For continuous outcomes, such a model specification is termed MMRM (mixed-model random missingness) by Mallinckrodt et al. (2001a, b). Thus, MMRM is a particular form of a linear mixed model, relevant for not only the clinical trials context, but also more generally for reasonably balanced studies with human subjects, and fitting within the ignorable likelihood paradigm.

Such a direct-likelihood approach is a very promising alternative for the often used simple methods described in Sect. 34.4.

34.5.2 Weighted Generalized Estimating Equations

For non-Gaussian outcomes in the marginal model framework, Liang and Zeger (1986) pointed out that inferences based on GEE, using the observed data are valid only under the strong assumption that the underlying missing-data process is MCAR. This is due to the fact that GEE is frequentist rather than likelihood based in nature. An exception is the situation where the working correlation structure happens to be correct, since then the estimating equations can be interpreted as likelihood equations. In general, however, the working correlation structure will not be correctly specified and hence, Robins et al. (1995) proposed a class of weighted estimating equations to allow for data which are MAR in case missingness is due to dropout.

The idea of weighted estimating equations (WGEE) is to weigh each subject's contribution in the GEEs by the inverse probability that a subject drops out at the time he or she dropped out. The incorporation of these weights reduces possible bias in the regression parameter estimates, $\hat{\vec{\beta}}$. Such a weight can be expressed as

$$\nu_{ij} \equiv P[D_i = j] = \prod_{k=2}^{j-1} (1 - P\left[R_{ik} = 0 | R_{i2} = \ldots = R_{i,k-1} = 1\right]) \\ \times P\left[R_{ij} = 0 | R_{i2} = \ldots = R_{i,j-1} = 1\right]^{I\{j \leq T\}},$$

where $j = 2, 3, \ldots, n+1$. Recall that we partitioned $\vec{Y}_i$ into the unobserved components $\vec{Y}_i^m$ and the observed components $\vec{Y}_i^o$. Similarly, we can make the same partition of $\vec{\mu}_i$ into $\vec{\mu}_i^m$ and $\vec{\mu}_i^o$. In the weighted GEE approach, the score equations to be solved when taking into account the correlation structure are:

$$S(\vec{\beta}) = \sum_{i=1}^{N} \frac{1}{\nu_{id}} \frac{\partial \vec{\mu}_i}{\partial \vec{\beta}'} \left(A_i^{1/2} R_i A_i^{1/2}\right)^{-1} \left(\vec{y}_i - \vec{\mu}_i\right) = \vec{0}, \tag{34.21}$$

where $\vec{y}_i(d)$ and $\vec{\mu}_i(d)$ are the first $d-1$ elements of $\vec{y}_i$ and $\vec{\mu}_i$ respectively. We define $\frac{\partial \vec{\mu}_i}{\partial \vec{\beta}'}(d)$ and $(A_i^{1/2} R_i A_i^{1/2})^{-1}(d)$ analogously, in line with the definitions of Robins et al. (1995).

34.5.3 Multiple Imputation

A well-known alternative method valid under the MAR assumption is multiple imputation (MI), which was introduced by Rubin (1978) and further discussed in Schafer (1999) and Little and Rubin (2002). The key idea of the procedure is to first replace each missing value with a set of M plausible values drawn from the conditional distribution of the unobserved outcomes, given the observed ones. This conditional distribution represents the uncertainty about the right value to impute. In this way, M imputed data sets are generated (imputation stage), which are then analyzed using standard complete data methods (analysis stage). Finally, the results from the M analyses have to be combined into a single one (pooling stage) by means of the method laid out in Rubin (1978), to produce inferences. To be valid, multiple imputation in its basic form requires the missingness mechanism to be MAR, even though versions under MNAR have been proposed (Rubin 1987; Molenberghs et al. 1997).

Thus, multiple imputation is valid under the same conditions as direct-likelihood, and therefore does not suffer from the problems encountered in most single imputation methods. However, there are a number of situations where multiple imputation is particularly useful. For example, when outcomes as well as covariates are missing then multiple imputation is a sensible route. The method is also useful when several analysis, perhaps conducted by different analysts, have to be done on the same set of incomplete data. In such a case, all analyses could start from the same set of multiply imputed sets of data and enhance comparability.

Let us discuss multiple imputation in a little more detail. In line with the notation in Sect. 34.3, suppose the parameter vector of the distribution of the response $\vec{Y}_i = (\vec{Y}_i^o, \vec{Y}_i^m)$ is denoted by $\vec{\theta} = (\vec{\beta}, \vec{\alpha})'$, in which $\vec{\beta}$ denotes the vector of fixed-effects parameters and $\vec{\alpha}$ the vector of covariance parameters. Multiple imputation uses the observed data $\vec{Y}^o$ to estimate the conditional distribution of $\vec{Y}^m$ given $\vec{Y}^o$. The missing data are sampled several times from this conditional distribution and augmented to the observed data. The resulting completed data are then used to estimate $\vec{\theta}$. If the distribution of $\vec{Y}_i = (\vec{Y}_i^o, \vec{Y}_i^m)$ were known, with parameter vector $\vec{\theta}$, then $\vec{Y}_i^m$ could be imputed by drawing a value of $\vec{Y}_i^m$ from the conditional distribution $f(\vec{y}_i^m | \vec{y}_i^o, \vec{\theta})$. The objective of the imputation phase is to sample from this true predictive distribution. However, $\vec{\theta}$ in the imputation model is unknown, and therefore needs to be estimated from the data first, say $\widehat{\vec{\theta}}$, after which $f(\vec{y}_i^m | \vec{y}_i^o, \widehat{\vec{\theta}})$ is used to impute the missing data. Precisely, this implies one first generates draws from the distribution of $\widehat{\vec{\theta}}$, thereby taking sampling uncertainty into account. Generally, the parameter vector in the imputation model differs from the parameter vector that governs the analysis model. Alternatively, a Bayesian approach, in which uncertainty about $\vec{\theta}$ is incorporated by means of some prior distribution for $\vec{\theta}$, can also be taken.

In the last phase of multiple imputation, the results of the analysis for the M imputed data sets are pooled into a single inference. The combined point estimate for the parameter of interest from the multiple imputation is simply the average of the M complete-data point estimates (Schafer 1999). Let $\vec{\theta}$ denote the parameter of interest, then the estimate and its estimated variance are given by:

$$\widehat{\vec{\theta}} \equiv \frac{1}{M}\sum_{m=1}^{M}\widehat{\vec{\theta}}^{m} \quad \text{and} \quad \widehat{\text{Var}}(\widehat{\vec{\theta}}) \equiv \vec{V} = \vec{W} + \left(\frac{M+1}{M}\right)\vec{B},$$

where

$$\vec{W} = \frac{\sum_{m=1}^{M}\vec{U}^{\,m}}{M} \quad \text{and} \quad \vec{B} = \frac{\sum_{m=1}^{M}(\widehat{\vec{\theta}}^{m} - \widehat{\vec{\theta}})(\widehat{\vec{\theta}}^{m} - \widehat{\vec{\theta}})'}{M-1},$$

with $\vec{W}$ denoting the average *within* imputation variance and $\vec{B}$ the *between* imputation variance (Rubin 1987).

As mentioned earlier, one can choose any complete data method to analyze the imputed data sets in the analysis stage of the MI method. This is particularly useful when interest lies in a marginal model for non-Gaussian outcomes. WGEE has been proposed as an alternative for GEE to be able to obtain valid inferences based on incomplete data under the MAR assumption. However, in WGEE, all subjects are given weights, which are calculated using the hypothesized dropout model. As a consequence, any misspecification of this dropout model will affect all subjects and thus all results. Therefore, we can consider MI combined with GEE at the analysis phase (MI-GEE). In essence, this method comes down to first using the predictive distribution of the unobserved outcomes given the observed ones and perhaps covariates. After this step, the missingness mechanism can be further ignored, provided we assume MAR. In these MI cases, a misspecification made in the imputation step will only effect the unobserved (i.e., imputed) but not the observed part of the data. Meng's (1994) results show that, as long as the imputation model is not grossly misspecified, this approach will perform well. Moreover, Beunckens et al. (2008) have proven that, even though WGEE is asymptotically unbiased, MI-GEE is less biased and more precise in small and moderate samples and MI-GEE is more robust to misspecification in either the imputation or measurement model.

34.6 Simple Methods and MAR Methods Applied to the ARMD Trial

Let us now apply the most often used simple methods which require the missingness mechanism to be MCAR, complete case analysis and last observation carried forward, as well as the following MAR methods to the age-related macular degeneration trial: direct-likelihood for Gaussian outcomes and when considering a random-effects model for non-Gaussian outcomes, and WGEE and MI-GEE when focus is on the marginal approach for non-Gaussian measurements.

Recall that there are 240 subjects, 188 of which have a complete follow-up. Note that of the 52 subjects with incomplete follow-up, 9 exhibit a non-monotone pattern. While this does not hamper direct-likelihood analysis, it is a challenge for WGEE. One way forward is to monotonize the missingness patterns by means of multiple imputation and then conduct WGEE, or to switch to MI-GEE altogether. However, to be consistent throughout the analysis, those subjects with non-monotone profiles will not be taken into account. Further, neither will the subjects without any follow-up measurement be included in the analysis, since at least one measurement should be taken. This results in a data set of 226 of 240 subjects, or 94.17%.

The original outcome is the visual acuity, that is the number of letters correctly read on a vision chart, which will be considered continuous. The dichotomous outcome is defined as increase or decrease in number of letters read compared with baseline. Both outcomes will be analyzed using the methods mentioned in the following text.

34.6.1 Analysis of the Continuous Outcome

We consider a multivariate normal model, with unconstrained time trend under placebo, a time-specific treatment effect, and an unstructured variance-covariance matrix. Let Y_{ij} be the visual acuity of subject $i = 1, \dots, 226$, at time point $j = 1, \dots, 4$, and T_i the treatment assignment for subject i, then the mean model takes the following form:

$$E(Y_{ij}) = \beta_{j1} + \beta_{j2} T_i.$$

Thus, this longitudinal model features a full treatment by time interaction with eight mean model parameters. The direct-likelihood analysis based on all observed data is contrasted with the simple CC and LOCF analyses.

Results of these three analyses are displayed in Table 34.4. From the parameter estimates, it is clear that the treatment effects are underestimated when considering the completers only. Whereas for all observed data treatment effect at week 12 and 52 are borderline significant, both turn insignificant when deleting subjects with missing values. For the LOCF analysis, going from week 4 to the end of the study, the underestimation of the treatment effect increases. Therefore, the effect at week 12 is borderline significant, but at week 52 it becomes insignificant. Once again, CC and LOCF miss important treatment differences, the most important one being that at week 52, the end of the study.

34.6.2 Analysis of the Binary Outcome

We now switch to the binary outcome, which indicates whether the number of letters correctly read at the follow-up occasion is higher or lower than the corresponding number of letters at baseline. Both marginal models and random-effects models are considered.

Table 34.4 Age-related macular degeneration trial. Parameter estimates (standard errors) for the linear mixed models, fitted to the continuous outcome visual acuity on the CC and LOCF population, and on the observed data (direct-likelihood). p-values are given for treatment effect at each of the four time points

Effect	Parameter	CC		LOCF		Observed data	
Parameter estimates (standard errors)							
Intercept 4	β_{11}	54.47	(1.54)	54.00	(1.47)	54.00	(1.47)
Intercept 12	β_{21}	53.08	(1.66)	53.03	(1.59)	53.01	(1.60)
Intercept 24	β_{31}	49.79	(1.80)	49.35	(1.72)	49.20	(1.74)
Intercept 52	β_{41}	44.43	(1.83)	44.59	(1.74)	43.99	(1.79)
Treatment effect 4	β_{12}	−2.87	(2.28)	−3.11	(2.10)	−3.11	(2.10)
Treatment effect 12	β_{22}	−2.89	(2.46)	−4.45	(2.27)	−4.54	(2.29)
Treatment effect 24	β_{32}	−3.27	(2.66)	−3.41	(2.45)	−3.60	(2.49)
Treatment effect 52	β_{42}	−4.71	(2.70)	−3.92	(2.48)	−5.18	(2.59)
p-values							
Treatment effect 4	β_{12}		0.211		0.140		0.140
Treatment effect 12	β_{22}		0.241		0.051		0.048
Treatment effect 24	β_{32}		0.220		0.165		0.150
Treatment effect 52	β_{42}		0.083		0.115		0.046

34.6.2.1 Marginal Models

In this section, a marginal model is used and in line with the previous section, we compare analyses performed on the completers only (CC), on the LOCF imputed data, as well as on the observed data. In all cases, standard GEE will be considered. For the observed, partially incomplete data, GEE is supplemented with WGEE and MI-GEE. Results of the GEE analyses are reported in Table 34.5. In all cases, we use the logit link, and the model takes the form

$$\text{logit}[P(Y_{ij} = 1 \mid T_i, t_j)] = \beta_{j1} + \beta_{j2} T_i, \tag{34.22}$$

with notational conventions as before, except that Y_{ij} is the indicator for whether or not letters of vision have been lost for subject i at time j, relative to baseline.

A working exchangeable correlation matrix is considered. For the WGEE analysis, the following weight model is assumed:

$$\begin{aligned} \text{logit}[P(D_i = j \mid D_i \geq j)] = \psi_0 &+ \psi_1 y_{i,j-1} + \psi_2 T_i \\ &+ \psi_{3,1} L_{1i} + \psi_{3,2} L_{2i} + \psi_{3,3} L_{3i} \\ &+ \psi_{4,1} I(t_j = 2) + \psi_{4,2} I(t_j = 3), \end{aligned}$$

where $y_{i,j-1}$ is the binary outcome at the previous time $t_{i,j-1} = t_{j-1}$, $L_{ki} = 1$ if the patient's eye lesion is of level $k = 1, \ldots, 4$ (since one dummy variable is redundant, only three are used), and $I(\cdot)$ is the indicator function. Parameter estimates, standard errors and p-values for the dropout model are given in Table 34.6. Covariates of importance are treatment assignment, the level of lesions at baseline, and time at which dropout occurs. For the latter covariates, there are three levels, since dropout can occur at times 2, 3, or 4. Hence, two indicator variables are included. Finally, the previous outcome does not have a significant impact, but will be kept in the model nevertheless.

When comparing parameter estimates across CC, LOCF, and observed data analyses, it is clear that LOCF has the effect of artificially increasing the correlation between measurements. The effect is mild in this case. The parameter estimates of the observed-data GEE are close to the LOCF results for earlier time points and close to CC for later time points. This is to be expected, as at the start of the study the LOCF and observed populations are virtually the same, with the same holding between CC and observed populations near the end of the study. Note also that the treatment effect under LOCF, especially at 12 weeks and after 1 year, is biased downward in comparison to the GEE analyses. Next, to properly use the information in the missingness process, WGEE or MI-GEE can be used. Two versions of MI-GEE are considered, that is, first the continuous outcome defined by the difference in numbers of letters correctly read compared with baseline is imputed whereafter the dichotomized version is analyzed, and second the binary outcome is imputed and analyzed.

In spite of there being no strong evidence for MAR, the results between GEE and WGEE differ quite a bit. It is noteworthy that at 12 weeks, a treatment effect

Table 34.5 Age-related macular degeneration trial. Parameter estimates (empirically corrected standard errors) for the marginal models: standard GEE on the CC and LOCF population and on the observed data. In the latter case, standard GEE, WGEE, and MI-GEE with imputation based on the continuous and binary outcome is used

				Observed data			
						Continuous	Binary
Effect	Parameter	CC	LOCF	Unweighted	WGEE	MI-GEE	MI-GEE
Parameter estimates (standard errors)							
Intercept 4	β_{11}	−1.01 (0.24)	−0.95 (0.21)	−0.95 (0.21)	−0.98 (0.44)	−0.95 (0.21)	−0.95 (0.21)
Intercept 12	β_{21}	−0.89 (0.24)	−0.99 (0.21)	−1.01 (0.22)	−1.77 (0.37)	−1.00 (0.22)	−0.98 (0.22)
Intercept 24	β_{31}	−1.13 (0.25)	−1.09 (0.22)	−1.07 (0.23)	−1.11 (0.33)	−1.05 (0.22)	−1.06 (0.25)
Intercept 52	β_{41}	−1.64 (0.29)	−1.46 (0.24)	−1.64 (0.29)	−1.72 (0.39)	−1.52 (0.26)	−1.57 (0.29)
Treatment 4	β_{12}	0.40 (0.32)	0.32 (0.29)	0.32 (0.29)	0.78 (0.66)	0.32 (0.29)	0.32 (0.29)
Treatment 12	β_{22}	0.49 (0.31)	0.59 (0.29)	0.62 (0.29)	1.83 (0.60)	0.60 (0.29)	0.58 (0.29)
Treatment 24	β_{32}	0.48 (0.33)	0.46 (0.29)	0.43 (0.30)	0.72 (0.53)	0.40 (0.30)	0.42 (0.32)
Treatment 52	β_{42}	0.40 (0.38)	0.32 (0.33)	0.40 (0.37)	0.72 (0.52)	0.33 (0.35)	0.31 (0.41)
Corr.	ρ	0.39	0.44	0.39	0.33	0.39	0.38
p-values							
Treatment 4	β_{12}	0.209	0.268	0.268	0.242	0.268	0.268
Treatment 12	β_{22}	0.113	0.040	0.034	0.003	0.037	0.048
Treatment 24	β_{32}	0.141	0.119	0.151	0.176	0.182	0.195
Treatment 52	β_{42}	0.283	0.323	0.277	0.162	0.349	0.456

Table 34.6 Age-related macular degeneration trial. Parameter estimates (standard errors) and p-values for a logistic regression model to describe dropout

Effect	Parameter	Estimate (s.e.)	p-value
Intercept	ψ_0	0.13 (0.49)	0.7930
Previous outcome	ψ_1	0.04 (0.38)	0.9062
Treatment	ψ_2	−0.87 (0.37)	0.0185
Lesion level 1	ψ_{31}	−1.82 (0.49)	0.0002
Lesion level 2	ψ_{32}	−1.89 (0.52)	0.0003
Lesion level 3	ψ_{33}	−2.79 (0.72)	0.0001
Time 2	ψ_{41}	−1.73 (0.49)	0.0004
Time 3	ψ_{42}	−1.36 (0.44)	0.0019

is observed with WGEE which is borderline with the other marginal analyses. However, as mentioned before, the beneficial property of unbiasedness for WGEE is merely fulfilled for very large samples. On the other hand, MI-GEE produces only a small amount of bias in small samples, which is less compared to the bias of WGEE. Moreover, MI-GEE is robust against misspecification of the imputation and measurement model. Therefore, in real-life settings such as the age-related macular degeneration trial, we would opt to use MI-GEE instead of WGEE. Both versions of MI-GEE considered here show quite similar results. The standard errors are smaller compared to the ones estimated using WGEE. The treatment effect at week 12 detected by WGEE becomes borderline again in both MI-GEE analyses. Further, also the p-values at later time points are larger compared to WGEE.

34.6.2.2 Random-Effects Models

Let us now turn to a random-intercepts logistic model, in spirit to Eq. 34.22:

$$\text{logit}[P(Y_{ij} = 1 \mid T_i, t_j)] = \beta_{j1} + b_i + \beta_{j2} T_i, \tag{34.23}$$

with notation as before and $b_i \sim N(0, \tau^2)$. For the model fitting, numerical integration is used. Results are shown in Table 34.7.

We observe the usual relationship between the marginal parameters of Table 34.5 and their random-effects counterparts. Note also that the random-intercepts variance is largest under LOCF, underscoring again that this method artificially increases the association between measurements on the same subject. In this case, in contrast to the marginal models, both LOCF and CC, considerably overestimate the treatment effect at certain times, in particular at 4 and 24 weeks (unlike the continuous case). In conclusion, it is clear that CC and LOCF may differ from the direct-likelihood and weighted or MI-based GEE analyses. This underscores that the latter analyses are to be considered as candidates for primary analysis.

Table 34.7 Age-related macular degeneration trial. Parameter estimates (standard errors) for the random-intercept models: numerical-integration based fits on the CC and LOCF population, and on the observed data (direct-likelihood)

Effect	Parameter	CC	LOCF	Direct-lik.
Intercept 4	β_{11}	−1.73 (0.42)	−1.76 (0.40)	−1.64 (0.37)
Intercept 12	β_{21}	−1.53 (0.41)	−1.85 (0.40)	−1.75 (0.38)
Intercept 24	β_{31}	−1.93 (0.43)	−2.01 (0.41)	−1.85 (0.39)
Intercept 52	β_{41}	−2.74 (0.48)	−2.66 (0.44)	−2.76 (0.47)
Treatment 4	β_{12}	0.64 (0.54)	0.55 (0.53)	0.51 (0.49)
Treatment 12	β_{22}	0.81 (0.53)	1.04 (0.53)	1.02 (0.50)
Treatment 24	β_{32}	0.77 (0.55)	0.80 (0.53)	0.70 (0.51)
Treatment 52	β_{42}	0.60 (0.59)	0.54 (0.56)	0.61 (0.59)
Random-intercept s.d.	τ	2.19 (0.27)	2.46 (0.27)	2.21 (0.26)
Random-intercept var.	τ^2	4.80 (1.17)	6.03 (1.33)	4.90 (1.14)

34.7 From MAR to Sensitivity Analysis

We have seen from previous sections that, if the MAR assumption is guaranteed to hold, a standard analysis will follow. This is certainly true for likelihood methods. Indeed, for example, a linear mixed model or a generalized linear mixed model fitted to incomplete data is valid, and it is as simple to conduct as it would be in contexts where data are complete.

The situation is a little different for the marginal GEE methods, since they need to be adjusted to the MAR case. Here, a mild extension of GEE to weighted GEE or a combination with multiple imputation comes to the rescue.

In conclusion, there are a number of tools available for correlated data which are easy to conduct and valid in the important MAR setting. As a consequence, there is little or no need for the simple methods such as complete case analysis or LOCF.

On the other hand though, in realistic settings, the reasons for missingness are varied and it is therefore hard to fully justify the MAR assumption as such. Moreover, since it is not possible to test for MNAR against MAR (Jansen et al. 2006; Molenberghs and Kenward 2007), one can never rule out the possibility of missing data to be MNAR. This implies that the need may exist to consider MNAR models. Nevertheless, ignorable analyses may provide reasonable stable results, even when the assumption of MAR is violated, in the sense that such analyses constrain the behavior of unseen data to be similar to that of the observed data (Mallinckrodt et al. 2001a, b).

While MNAR models are more general and explicitly incorporate the dropout mechanism, the inferences they produce are typically highly dependent on untestable and often implicit assumptions regarding the distribution of the unobserved measurements given the observed ones. The quality of the fit to the observed data does not reflect at all the appropriateness of the implied structure governing the unobserved data. This implies that a definite MNAR analysis does not exist.

Therefore, a sensible compromise between blindly shifting to MNAR models or ignoring them altogether is to make them a component of a sensitivity analysis. In this way we can explore the impact of deviations from the MAR assumption on the conclusions. In that sense, it is important to consider the effect on key parameters.

When the missingness mechanism is (assumed to be) MNAR, all frameworks have their advantages and disadvantages. However, in the case of pattern-mixture modeling it is clear which parts of the model are identifiable from the data, and which parts are completely assumption driven. On the other hand, selection models may be easier to conduct randomization-consistent inferences.

In this section we consider the full selection model proposed by Diggle and Kenward (1994). Next, an overview of different strategies to fit pattern-mixture models is provided. One route for sensitivity analysis is to explore the effect of different plausible models within one modeling framework. Another possibility is to consider pattern-mixture models as a complement to selection models (Thijs et al. 2002; Michiels et al. 2002). Finally, we discuss local influence, a sensitivity tool to explore the effect of (small) changes from the posited MAR model within the selection model framework.

34.7.1 Selection Models

Much of the early development of, and debate about, selection models appeared in the econometrics literature in which the tobit model (Heckman 1976) played a central role. This combines a marginal Gaussian regression model for the response, as might be used in the absence of missing data, with a Gaussian-based threshold model for the probability of a value being missing. Later on, selection models have been proposed for longitudinal data in the biometric and epidemiological setting.

For continuous outcomes, Diggle and Kenward (1994) proposed a full selection model which is valid under MNAR. In the discrete case, Molenberghs et al. (1997) considered a global odds ratio (Dale) model. Within the selection model framework, models have been proposed for non-monotone missingness as well (Baker et al. 1992; Jansen and Molenberghs 2008), and further a number of proposals have been made for non-Gaussian outcomes (Molenberghs and Verbeke 2005). Let us describe the Diggle and Kenward model for continuous longitudinal data in more detail, since we will apply this to the ARMD trial later on.

Diggle and Kenward Model for Continuous Longitudinal Data Diggle and Kenward (1994) proposed a model for longitudinal Gaussian data with non-random dropout, that is, the missingness mechanism was assumed to be MNAR, which combines the multivariate normal model for longitudinal Gaussian data with a logistic regression for dropout. To maximize the resulting likelihood, integration over the missing data is needed.

The likelihood contribution of the ith subject, based on the observed data $(\vec{y}_i^o, d_i)$, is proportional to the marginal density function

$$\begin{aligned} f(\vec{y}_i^o, d_i|\vec{\theta}, \vec{\psi}) &= \int f(\vec{y}_i, d_i|\vec{\theta}, \vec{\psi})\, d\vec{y}_i^m \\ &= \int f(\vec{y}_i|\vec{\theta}) f(d_i|\vec{y}_i, \vec{\psi})\, d\vec{y}_i^m, \end{aligned} \tag{34.24}$$

in which a marginal model for $\vec{Y}_i$ is combined with a model for the dropout process, conditional on the response – since we are considering the selection model framework – and where $\vec{\theta}$ and $\vec{\psi}$ are vectors of unknown parameters in the measurement model and dropout model, respectively.

Let $\vec{h}_{ij} = (y_{i1}, \ldots, y_{i,j-1})$ denote the observed history of subject i up to time $t_{i,j-1}$. The Diggle–Kenward model for the dropout process allows the conditional probability for dropout at occasion j, given that the subject was still observed at the previous occasion, to depend on the history $\vec{h}_{ij}$ and the possibly unobserved current outcome y_{ij}, but not on future outcomes $y_{ik}, k > j$. These conditional probabilities $P(D_i = j|D_i \geq j, \vec{h}_{ij}, y_{ij}, \vec{\psi})$ can now be used to calculate the probability of dropout at each occasion:

$$f(d_i|\vec{y}_i, \vec{\psi}) = P(D_i = d_i|\vec{y}_i, \vec{\psi}) = P(D_i = d_i|\vec{h}_{id_i}, y_{id_i}, \vec{\psi}) \tag{34.25}$$

$$= \begin{cases} P(D_i = d_i|D_i \geq d_i, \vec{h}_{id_i}, y_{id_i}, \vec{\psi}), & d_i = 2, \\ P(D_i = d_i|D_i \geq d_i, \vec{h}_{id_i}, y_{id_i}, \vec{\psi}) & \\ \qquad \times \prod_{j=2}^{d_i-1} \left[1 - P(D_i = j|D_i \geq j, \vec{h}_{ij}, y_{ij}, \vec{\psi})\right], & d_i = 3, \ldots, n, \\ \prod_{j=2}^{n} \left[1 - P(D_i = j|D_i \geq j, \vec{h}_{ij}, y_{ij}, \vec{\psi})\right], & d_i = n+1. \end{cases}$$

Diggle and Kenward (1994) combine a multivariate normal model for the measurement process with a logistic regression model for the dropout process. More specifically, the measurement model assumes that the vector $\vec{Y}_i$ of repeated measurements for the ith subject satisfies the linear regression model

$$\vec{Y}_i \sim N(\boldsymbol{X}_i\vec{\beta}, V_i), \qquad i = 1, \ldots, N, \tag{34.26}$$

in which $\vec{\beta}$ is a vector of population-averaged regression coefficients. The matrix V_i can be left unstructured or assumed to be of a specific form, such as resulting from a linear mixed model.

The logistic dropout model can, for example, take the form

$$\text{logit}\left[P(D_i = j \mid D_i \geq j, \vec{h}_{ij}, y_{ij}, \vec{\psi})\right] = \psi_0 + \psi_1 y_{i,j-1} + \psi_2 y_{ij}. \tag{34.27}$$

More general models can easily be constructed by including the complete history $\vec{h}_{ij} = (y_{i1}, \ldots, y_{i,j-1})$, as well as external covariates, in the above conditional dropout model. Note also that, strictly speaking, one could allow dropout at a specific occasion to be related to all future responses as well. However, this is rather counterintuitive in many cases. Moreover, including future outcomes seriously complicates the calculations since computation of the likelihood (Eq. 34.24) then requires evaluation of a possibly high-dimensional integral. Note also that special cases of the model (34.27) are obtained from setting $\psi_2 = 0$ or $\psi_1 = \psi_2 = 0$, respectively. In the first case, dropout is no longer allowed to depend on the current measurement, implying MAR. In the second case, dropout is independent of the outcome, which corresponds to MCAR. In both cases, all parameters can be estimated using standard software since the multivariate normal measurement model and the dropout model can then be fitted separately.

Diggle and Kenward (1994) obtained parameter and precision estimates by maximum likelihood. The likelihood involves marginalization over the unobserved outcomes $\vec{Y}_i^m$, for which subject-by-subject integration is required. Practically, this involves relatively tedious and computationally demanding forms of numerical integration. Diggle and Kenward (1994) used the Nelder and Mead simplex algorithm (Nelder and Mead 1965), whereas we will use the Newton-Raphson Ridge optimization method.

34.7.2 Selection Models Applied to the ARMD Trial

In this section, the visual acuity is first analyzed using the full selection model proposed by Diggle and Kenward (1994), discussed in Sect. 34.7.1. Apart from modeling the three missing data mechanisms MCAR, MAR, and MNAR, explicitly, an ignorable MAR analysis is also conducted in which the model for the response measurements only was fitted. For the measurement model, the linear mixed model was used, assuming again different intercepts and treatment effects for each of the four time points, next to an unstructured variance-covariance matrix. In the full selection models, the dropout is modeled by Eq. 34.27. Parameter estimates and corresponding standard errors of the fixed effects of the measurement model and of the dropout model parameters are given in Table 34.8. As expected, the parameter estimates and standard errors coincide for the ignorable direct-likelihood analysis and the selection models under MCAR and MAR, except for some numerical noise.

Since the main interest lies in the treatment effect at 1 year, the corresponding p-values are displayed in Table 34.8. In all four cases, this treatment effect is (borderline) significant.

Table 34.8 Age-related macular degeneration trial. Parameter estimates (standard errors) assuming ignorability, as well as explicitly modeling the missing data mechanism under MCAR, MAR, and MNAR assumptions, for all data

All subjects		Ignorable		MCAR		MAR		MNAR	
Effect	Parameters	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)
Measurement model									
Intercept 4	β_{11}	54.00	(1.47)	54.00	(1.46)	54.00	(1.47)	54.00	(1.47)
Intercept 12	β_{21}	53.01	(1.60)	53.01	(1.59)	53.01	(1.60)	52.98	(1.60)
Intercept 24	β_{31}	49.20	(1.74)	49.20	(1.73)	49.19	(1.74)	49.06	(1.74)
Intercept 52	β_{41}	43.99	(1.79)	43.99	(1.78)	43.99	(1.79)	43.52	(1.82)
Treatment 4	β_{12}	−3.11	(2.10)	−3.11	(2.07)	−3.11	(2.09)	−3.11	(2.10)
Treatment 12	β_{22}	−4.54	(2.29)	−4.54	(2.25)	−4.54	(2.29)	−4.67	(2.29)
Treatment 24	β_{32}	−3.60	(2.49)	−3.60	(2.46)	−3.60	(2.50)	−3.80	(2.50)
Treatment 52	β_{42}	−5.18	(2.59)	−5.18	(2.57)	−5.18	(2.62)	−5.71	(2.63)
Dropout model									
Intercept	ψ_0			−2.79	(0.17)	−1.86	(0.46)	−1.81	(0.47)
Previous	ψ_1					−0.020	(0.009)	0.016	(0.022)
Current	ψ_2							−0.042	(0.023)
−2 log-likelihood		6,488.7		6,782.7		6,778.4		6,775.9	
Treatment effect at 1 year	p-value	0.046		0.044		0.048		0.030	

Note that for the MNAR analysis, the estimates of the ψ_1 and ψ_2 parameter are more or less of the same magnitude, but with a different sign. This is in line with the argument of Molenberghs et al. (2001), stating that the dropout oftentimes depends on the increment $y_{ij} - y_{i,j-1}$. This is because two subsequent measurements are usually positively correlated. By rewriting the fitted dropout model in terms of the increment,

$$\text{logit}\left[\text{pr}(D_i = j | D_i \geq j, \vec{h}_{ij}, y_{ij}, \vec{\psi})\right] = -1.81 - 0.026 y_{i,j-1} - 0.042(y_{ij} - y_{i,j-1}),$$

we find that the probability of dropout increases with larger negative increments; that is, those patients who showed or would have shown a greater decrease in visual acuity from the previous visit are more likely to drop out.

34.7.3 Pattern-Mixture Models

Pattern-mixture models were introduced in Sect. 34.3.1 as one of the three major frameworks within which missing data models can be developed. In this section we provide a brief overview of such pattern-mixture models. More details can be found in Verbeke and Molenberghs (2000) and Molenberghs and Kenward (2007). Early references include Rubin (1978), who mentioned the concept of a sensitivity analysis within a fully Bayesian framework, Glynn et al. (1986), and Little and Rubin (1987). Important early development was provided by Little (1993, 1994, 1995).

Pattern-mixture models can be considered for their own sake to answer a particular scientific question. Further, several authors have contrasted selection models and pattern-mixture models. This is done either (1) to answer the same scientific question, such as marginal treatment effect or time evolution, based on these two rather different modeling strategies, or (2) to gain additional insight by supplementing the selection model results with those from a pattern-mixture approach.

Examples of pattern-mixture applications can be found in Verbeke et al. (2001a) or Michiels et al. (2002) for continuous outcomes, and Molenberghs et al. (1999) or Michiels et al. (1999) for categorical outcomes.

Recall that the pattern-mixture decomposition is given by Eq. 34.5. As a simple illustration consider a continuous response at three times of measurement which will be modeled using a trivariate Gaussian distribution. Assume that there may be dropout at time 2 or 3, and let the dropout indicator R take the values 1 and 2 to indicate that the last observation occurred at these times and 3 to indicate no dropout. Then, in the first instance, the model implies a different distribution for each time of dropout. We can write:

$$\vec{y} \mid r \sim N\left(\vec{\mu}^{(r)}; \vec{\Sigma}^{(r)}\right), \tag{34.28}$$

where

$$\vec{\mu}^{(r)} = \begin{bmatrix} \mu_1^{(r)} \\ \mu_2^{(r)} \\ \mu_3^{(r)} \end{bmatrix} \quad \text{and} \quad \vec{\Sigma}^{(r)} = \begin{bmatrix} \sigma_{11}^{(r)} & \sigma_{21}^{(r)} & \sigma_{31}^{(r)} \\ \sigma_{21}^{(r)} & \sigma_{22}^{(r)} & \sigma_{32}^{(r)} \\ \sigma_{31}^{(r)} & \sigma_{32}^{(r)} & \sigma_{33}^{(r)} \end{bmatrix},$$

for $r = 1, 2, 3$. Let $\mathrm{P}(r) = \pi_r$, then the marginal distribution of the response is a mixture of normals with, for example, mean

$$\vec{\mu} = \sum_{r=1}^{3} \pi_r \vec{\mu}^{(r)}.$$

However, although the π_r can be simply estimated from the observed proportions in each dropout group, only 16 of the 27 response parameters can be identified from the data without making further assumptions. These 16 comprise all the parameters from the completers plus those from the following two submodels. For $r = 2$

$$N\left(\begin{bmatrix} \mu_1^{(2)} \\ \mu_2^{(2)} \end{bmatrix}; \begin{bmatrix} \sigma_{11}^{(2)} & \sigma_{21}^{(2)} \\ \sigma_{31}^{(3)} & \sigma_{32}^{(3)} \end{bmatrix}\right),$$

and for $r = 1$: $N\left(\mu_1^{(2)}; \sigma_{11}^{(1)}\right)$. This is a *saturated* pattern-mixture model and the representation makes it very clear what information each dropout group provides, and consequently the assumptions that need to be made if we are to predict the behavior of the unobserved responses, and so obtain marginal models for the response. If the three sets of parameters $\vec{\mu}^{(r)}$ are simply equated, this implies MCAR. Progress can be made with less stringent restrictions however. In practice, choice of restrictions will need to be guided by the context. In addition, the form of the data will typically be more complex, requiring, for example, a more structured model for the response with the incorporation of covariates. Hence such models can be constructed in many ways.

An important issue is that pattern-mixture models are by construction under-identified, that is, over-specified. Little (1993, 1994) solves this problem through the use of identifying restrictions: inestimable parameters of the incomplete patterns are set equal to (functions of) the parameters describing the distribution of the completers. Identifying restrictions are not the only way to overcome under-identification, and we will discuss alternative approaches. Although some authors perceive this under-identification as a drawback, it can be an asset because it forces one to reflect on the assumptions made. Pattern-mixture models can serve important roles in sensitivity analysis.

Fitting pattern-mixture models can be approached in several ways. It is important to decide whether pattern-mixture and selection models are to be contrasted with one another or rather the pattern-mixture modeling is the central focus. In the latter case, it is natural to conduct an analysis, and preferably a sensitivity analysis, *within* the pattern-mixture family. Basically we will consider three strategies to deal with under-identification.

- *Strategy 1.* As mentioned before, Little (1993, 1994) advocated the use of identifying restrictions and presented a number of examples. A general framework for identifying restrictions is discussed in more detail in Thijs et al. (2002), with three special but important cases: *complete case missing values* (CCMV) (proposed by Little 1993), *neighboring case missing values* (NCMV), and *available case missing values* (ACMV). Note that ACMV is the natural counterpart of MAR in the pattern-mixture model framework (Molenberghs et al. 1998). This provides a way to compare ignorable selection models with their counterpart in the pattern-mixture setting. Kenward et al. (2003) focus on restrictions avoiding dependence of dropout on measurements made at future occasions.

  The procedure to apply identifying restrictions is discussed in full detail in Thijs et al. (2002). The key steps are as follows:

  1. Fit a model to the pattern-specific identifiable densities: $f_t(y_1, \ldots, y_t)$. This results in a set of parameter estimates, $\vec{\beta}_p$ say, for each pattern p.
  2. Select an identification method of choice (ACMV, CCMV, NCMV).
  3. Using this identification method, determine the conditional distributions of the unobserved outcomes, given the observed ones:
  $$f_t(y_{t+1}, \ldots, y_t | y_1, \ldots, y_t).$$
  4. Using standard multiple imputation methodology (Rubin 1987; Schafer 1997; Verbeke and Molenberghs 2000; Molenberghs and Kenward 2007), draw multiple imputations for the unobserved components, given the observed outcomes and the correct pattern-specific density.
  5. Analyze the multiply imputed data sets using the method of choice. This can be another pattern-mixture model, but also a selection model or any other desired model.
  6. Inferences can be conducted in the standard multiple imputation way.

- *Strategy 2.* As opposed to identifying restrictions, model simplification can be done in order to identify the parameters. The advantage is that the number of parameters decreases, which is desirable since the length of the parameter vector is a general issue with pattern-mixture models. Indeed, Hogan and Laird (1997) noted that in order to estimate the large number of parameters in general pattern-mixture models, one has to make the awkward requirement that each dropout pattern occurs sufficiently often. Broadly, we distinguish between two types of simplifications.
  - *Strategy 2a.* Trends can be restricted to functional forms supported by the information available within a pattern. For example, a linear or quadratic time trend is easily extrapolated beyond the last obtained measurement. One only needs to provide an ad hoc solution for the first or the first few patterns. In order to fit such models, one simply has to carry out a model building exercise within each of the patterns separately.

- *Strategy 2b.* Next, one can let the parameters vary across patterns in a controlled parametric way. Thus, rather than estimating a separate time trend within each pattern, one could for example assume that the time evolution within a pattern is unstructured, but parallel across patterns. This is effectuated by treating pattern as a covariate. The available data can be used to assess whether such simplifications are supported within the time ranges for which there is information.

Although the second strategy is computationally simple, it is important to note that there is a price to pay. Indeed, simplified models, qualified as *assumption rich* by Sheiner et al. (1997), also make untestable assumptions, just as in the selection model case. From a technical point of view, Strategy 2 only requires to either consider "pattern" as an extra covariate in the model, or to conduct an analysis "by pattern," such that a separate analysis is obtained for each of the dropout patterns. In the identifying restrictions setting on the other hand (Strategy 1), the assumptions are clear from the start. Precisely for these reasons it is stated in Thijs et al. (2002) that the use of simplified models is not the best strategy and can be rather dangerous as well.

Pattern-mixture models do not always automatically provide estimates and standard errors of marginal quantities of interest, such as overall treatment effect or overall time trend. Hogan and Laird (1997) provided a way to derive selection model quantities from the pattern-mixture model. Several authors have followed this idea to formally compare the conclusions from a selection model with the selection model parameters in a pattern-mixture model (Michiels et al. 1999; Verbeke et al. 2001a).

34.7.4 Pattern-Mixture Models Applied to the ARMD Trial

In this section, we consider the use of pattern-mixture models for the visual acuity outcome. Since there are four fixed time points in the ARMD trial, there are four patterns, going from only one observation available (Pattern 1) to all planned measurements observed (Pattern 4). These four patterns have subsample sizes 188, 24, 8, and 6, respectively. The corresponding pattern probabilities are $\vec{\pi}' = (0.832, 0.106, 0.035, 0.027)'$.

34.7.4.1 Strategy 1

First, we will apply Strategy 1 making use of CCMV, NCMV, and ACMV identification restrictions. In order to apply such identifying restrictions, one first needs to fit a model to the observed data. We will opt for a saturated model with parameters specific to each pattern. Next, for each of the three restrictions, multiple imputation is conducted, after which the same multivariate normal model as before is fitted to the imputed datasets. Finally, results are combined into a single one. The results for the three types of restrictions are shown in Table 34.9. From the estimates and associated standard errors, it is clear that there is little difference in conclusions between the strategies. The estimates for treatment effect

Table 34.9 Age-related macular degeneration trial. Parameter estimates (standard errors) and *p*-values resulting from the pattern-mixture model using identifying restrictions ACMV, CCMV, and NCMV

		ACMV	CCMV	NCMV
Effect	Parameter			
Parameter estimate (standard error)				
Intercept 4	β_{11}	54.00 (1.47)	54.00 (1.47)	54.00 (1.47)
Intercept 12	β_{21}	53.22 (1.98)	52.89 (1.61)	52.97 (2.20)
Intercept 24	β_{31}	49.43 (2.14)	49.45 (1.79)	49.05 (2.49)
Intercept 52	β_{41}	44.73 (2.69)	44.67 (2.35)	44.40 (2.73)
Treatment 4	β_{12}	−3.11 (2.10)	−3.11 (2.10)	−3.11 (2.10)
Treatment 12	β_{22}	−4.94 (2.81)	−4.26 (2.36)	−4.56 (2.71)
Treatment 24	β_{32}	−4.21 (2.82)	−3.77 (2.55)	−3.79 (2.92)
Treatment 52	β_{42}	−5.19 (2.81)	−4.72 (2.60)	−4.76 (2.90)
p-value				
Intercept 4	β_{11}	<.0001	<.0001	<.0001
Intercept 12	β_{21}	<.0001	<.0001	<.0001
Intercept 24	β_{31}	<.0001	<.0001	<.0001
Intercept 52	β_{41}	<.0001	<.0001	<.0001
Treatment 4	β_{12}	0.140	0.140	0.140
Treatment 12	β_{22}	0.083	0.071	0.093
Treatment 24	β_{32}	0.139	0.140	0.200
Treatment 52	β_{42}	0.065	0.069	0.101

and corresponding standard errors obtained under CCMV and NCMV restrictions are underestimated when comparing to ACMV.

Further, we observe that for all three strategies the *p*-value for the treatment effect at 1 year is above the significance level of 0.05, yet it is borderline significant for the ACMV restrictions – which is equivalent to MAR – in line with the conclusions drawn from the selection models in the previous section. The *p*-value is closest to significance and thus to the one from the selection models in case the NCMV restrictions are considered. The Diggle–Kenward MAR and MNAR selection models also show a borderline significant treatment effect, its *p*-value being below 0.05.

34.7.4.2 Strategy 2

To apply Strategy 2a, we have to fit the posited measurement model within each of the four patterns separately. For consistency reasons, again a saturated model will be considered. In Strategy 2b the measurement model reflects dependence on dropout by allowing the fixed effects as well as the covariance parameters to change with dropout pattern. This is done by including the interaction of all effects in the model with the pattern variable. Since in Strategy 2a we fit the same saturated model for each pattern, the result will be the same as obtained by Strategy 2b.

As mentioned above, this second strategy is not the best one to choose (Molenberghs and Kenward 2007). The strategy uses models that are sufficiently

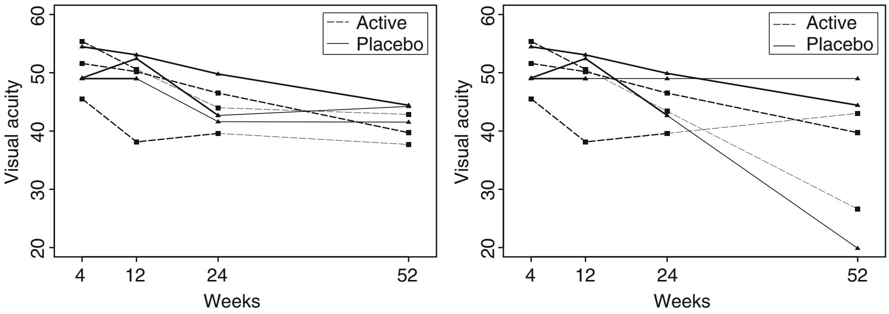


Fig. 34.6 Age-related macular degeneration trial. Mean profiles using ACMV restrictions (*left*) and extrapolation based on Strategy 2 (*right*), for dropout pattern 2–4, grouped per treatment arm

simple for extrapolations to be possible past the point of dropout within a particular pattern. Further, the results greatly depend on the extrapolations made, which may of course be highly inappropriate, especially as they are typically based on low-order polynomials. At the same time, the assumptions that these extrapolations imply concerning the dropout mechanisms are far from transparent. Let us illustrate this issue by comparing the extrapolations after applying Strategy 2 with the predicted means for each pattern obtained by the ACMV restriction from Strategy 1. We choose to consider the ACMV restriction due to its elegant property of being equivalent to MAR in the selection model framework. This is represented in Fig. 34.6. Bold line type is used for the range over which data are obtained for a particular pattern and extrapolation is indicated using thinner line type.

In all plots, the same mean response scale was retained, illustrating that the identifying restrictions strategies extrapolate much closer to the observed data mean response. Within both treatment arms there is a large difference between the extrapolation for pattern 2 and 3. The active treatment group shows a decline after dropout for pattern 2 in contrast to a rise after dropout for pattern 3. On the other hand, for the placebo group, pattern 2 remains stable after dropout, whereas pattern 3 shows a decrease after dropout. These findings suggest, again, that a more careful reflection on the extrapolation method is required.

34.7.5 Selection Models and Local Influence

Let us return to the Diggle and Kenward (1994) selection model, as described in Sect. 34.7.1, and consider the dropout model (34.27). When ω equals zero, the dropout model is random, and all parameters can be estimated using standard software since the measurement model for which we use a linear mixed model and the dropout model, assumed to follow a logistic regression, can then be fitted separately. If $\omega \neq 0$, the dropout process is assumed to be non-random.

Equation 34.6 is now used to construct the dropout process:

$$f(d_i|\vec{y}_i,\vec{\psi}) = \begin{cases} \prod_{j=2}^{n}\left[1-g\left(\vec{h}_{ij},y_{ij}\right)\right] & \text{for a complete sequence} \\ \prod_{j=2}^{d-1}\left[1-g\left(\vec{h}_{ij},y_{ij}\right)\right]g\left(\vec{h}_{id},y_{id}\right) & \text{for a dropout.} \end{cases} \tag{34.29}$$

Let us now shift attention to sensitivity and influence analysis issues. Whereas a global influence approach is based on case-deletion, a local influence-based sensitivity assessment of the relevant quantities, such as treatment effect or time evolution parameters, with respect to assumptions about the dropout model is based on the following perturbed version of Eq. 34.6:

$$\operatorname{logit}\left(g\left(\vec{h}_{ij},y_{ij}\right)\right) = \operatorname{logit}\left[\operatorname{pr}\left(D_i = j|D_i \geq j,\vec{y}_i\right)\right] = \vec{h}_{ij}\vec{\psi} + \omega_i y_{ij}, \tag{34.30}$$

$i = 1,\ldots,N$, in which different subjects give different weights to the response at time t_{ij} to predict dropout at time t_{ij}. If all ω_i equal zero, the model reduces to an MAR model. Hence Eq. 34.30 can be seen as an extension of the MAR model, which allows some individuals to drop out in a "less random" way ($|\omega_i|$ large) than others ($|\omega_i|$ small). It has to be noted that, even when ω_i is large, we still cannot conclude that the dropout model for these subjects is non-random. Rather, it is a way of pointing to subjects which, due to their strong influence, are able to distort the model parameters such that they can produce, for example, a dropout mechanism which is *seemingly* non-random. In reality, many different characteristics of such an individual's profile might be responsible for this effect. As mentioned earlier, such sensitivity has been alluded to by many authors, such as Laird (1994) and Rubin (1994).

Cook (1986) suggests that more confidence can be put in a model which is relatively stable under small modifications. The best known perturbation schemes are based on case-deletion (Cook and Weisberg 1982; Chatterjee and Hadi 1988) in which the effect is studied of completely removing cases from the analysis. They were introduced by Cook (1977, 1979) for the linear regression context. Denote the log-likelihood function, corresponding to measurement model (34.8) and dropout model (34.6), by

$$\ell(\vec{\gamma}) = \sum_{i=1}^{N}\ell_i(\vec{\gamma}), \tag{34.31}$$

in which $\ell_i(\vec{\gamma})$ is the contribution of the ith individual to the log-likelihood, and where $\vec{\gamma} = (\vec{\theta},\vec{\psi},\omega)$ is the s-dimensional vector, grouping the parameters of the measurement model and the dropout model. Further, we denote by

$$\ell_{(-i)}(\vec{\gamma}) \tag{34.32}$$

the log-likelihood function, where the contribution of the ith subject has been removed. Cook's distances are based on measuring the discrepancy between either the maximized likelihoods Eq. 34.31 and Eq. 34.32 or (subsets of) the estimated parameter vectors $\widehat{\vec{\gamma}}$ and $\widehat{\vec{\gamma}}_{(-i)}$, with obvious notation. Precisely, we will consider both

$$CD_{1i} = 2\left(\widehat{\ell} - \widehat{\ell}_{(-i)}\right) \tag{34.33}$$

as well as

$$CD_{2i}(\vec{\gamma}) = 2\left(\widehat{\vec{\gamma}} - \widehat{\vec{\gamma}}_{(-i)}\right)' \ddot{L}^{-1}\left(\widehat{\vec{\gamma}} - \widehat{\vec{\gamma}}_{(-i)}\right), \tag{34.34}$$

in which $\ddot{L}$ is the matrix of all second-order derivatives of $\ell(\vec{\gamma})$ with respect to $\vec{\gamma}$, evaluated at $\vec{\gamma} = \widehat{\vec{\gamma}}$. Formulation (34.34) easily allows to consider the global influence in a subvector of $\vec{\gamma}$, such as the dropout parameters $\vec{\psi}$, or the non-random parameter ω. This will be indicated using notation of the form $CD_{2i}(\vec{\psi})$, $CD_{2i}(\omega)$, etc.

In linear regression, global influence is conceptually simple, computationally straightforward, and well studied. The latter two of these features do not carry over to more general settings. To overcome these limitations, *local* influence methods have been suggested. The principle is to investigate how the results of an analysis are changed under infinitesimal perturbations of the model. In the framework of the linear mixed model, Beckman et al. (1987) used local influence to assess the effect of perturbing the error variances, the random-effects variances, and the response vector. In the same context, Lesaffre and Verbeke (1998) have shown that the local influence approach is also useful for the detection of influential subjects in a longitudinal data analysis. Moreover, since the resulting influence diagnostics can be expressed analytically, they often can be decomposed in interpretable components, which yield additional insights in the reasons why some subjects are more influential than others.

Verbeke et al. (2001a) studied the influence the non-randomness of dropout exerts on the model parameters. Let us briefly sketch the principles of local influence and then apply them to our MNAR problem.

We denote the log-likelihood function corresponding to model (34.30) by

$$\ell(\vec{\gamma}|\vec{\omega}) = \sum_{i=1}^{N} \ell_i(\vec{\gamma}|\omega_i), \tag{34.35}$$

in which $\ell_i(\vec{\gamma}|\omega_i)$ is the contribution of the ith individual to the log-likelihood, and where $\vec{\gamma} = (\vec{\theta}, \vec{\psi})$ is the s-dimensional vector, grouping the parameters of the measurement model and the dropout model, not including the $N \times 1$ vector $\vec{\omega} = (\omega_1, \omega_2, \ldots, \omega_N)'$ of weights defining the perturbation of the MAR model. Let $\widehat{\vec{\gamma}}$ be the maximum likelihood estimator for $\vec{\gamma}$, obtained by maximizing $\ell(\vec{\gamma}|\vec{\omega}_0)$, and let $\widehat{\vec{\gamma}}_\omega$ denote the maximum likelihood estimator for $\vec{\gamma}$ under $\ell(\vec{\gamma}|\vec{\omega})$. Cook (1986) proposed to measure the distance between $\widehat{\vec{\gamma}}_\omega$ and $\widehat{\vec{\gamma}}$ by the so-called likelihood displacement, defined by $LD(\vec{\omega}) = 2\left(\ell(\widehat{\vec{\gamma}}|\vec{\omega}_0) - \ell(\widehat{\vec{\gamma}}_\omega|\vec{\omega})\right)$. Since this quantity can only be depicted when $N = 2$, Cook (1986) proposed to look at local influence, i.e., at the normal curvatures $C_{\vec{h}}$ of $\vec{\xi}(\vec{\omega})$ in $\vec{\omega}_0$, in the direction of some N-dimensional vector $\vec{h}$ of unit length. It can be shown that a general form is given by

$$C_{\vec{h}}(\vec{\theta}) = -2\vec{h}'\left[\left.\frac{\partial^2 \ell_{i\omega}}{\partial\vec{\theta}\partial\omega_i}\right|_{\omega_i=0}\right]' \ddot{L}^{-1}(\vec{\theta})\left[\left.\frac{\partial^2 \ell_{i\omega}}{\partial\vec{\theta}\partial\omega_i}\right|_{\omega_i=0}\right]\vec{h}$$

$$C_{\vec{h}}(\vec{\psi}) = -2\vec{h}'\left[\left.\frac{\partial^2 \ell_{i\omega}}{\partial\vec{\psi}\partial\omega_i}\right|_{\omega_i=0}\right]' \ddot{L}^{-1}(\vec{\psi})\left[\left.\frac{\partial^2 \ell_{i\omega}}{\partial\vec{\psi}\partial\omega_i}\right|_{\omega_i=0}\right]\vec{h},$$

evaluated at $\vec{\gamma} = \widehat{\vec{\gamma}}$, where indeed the influence for the measurement and dropout model parameters split, since the second derivative matrix of the log-likelihood, $\ddot{L}$ is block-diagonal with blocks $\ddot{L}(\vec{\theta})$ and $\ddot{L}(\vec{\psi})$. Verbeke et al. (2001b) have decomposed local influence into meaningful and interpretable components.

34.7.6 Local Influence Applied to the ARMD Trial

Let us apply the local influence sensitivity tool to the ARMD trial. Figure 34.7 displays overall C_i and influences for subvectors $\vec{\theta}$, $\vec{\beta}$, $\vec{\alpha}$, and $\vec{\psi}$. In addition, the direction $\vec{h}_{\max}$, corresponding to maximal local influence, is given. The main emphasis should be put on the relative magnitudes. We observe that patients #10, #27, #28, #114, #139, and #154 have larger C_i values compared to other patients, which means they can be considered influential. Virtually the same picture holds for $C_i(\vec{\psi})$.

Turning attention to the influence on the measurement model, we see that for $C_i(\vec{\beta})$, there are no strikingly high peaks, whereas $C_i(\vec{\alpha})$ reveals two considerable peaks for patients #68 and #185. Note that both patients fail to have a high peak for the overall C_i. This is due to the fact that the scale for $C_i(\vec{\alpha})$ is relatively small, comparing to the overall C_i. Nevertheless, these patients can still be considered influential. Finally, the direction of maximum curvature reveals the same six influential patients as the overall C_i.

In Fig. 34.8, the individual profiles of the influential observations are highlighted. Let us take a closer look at these cases. The six patients, which are influencing the dropout model parameters, are those that drop out after the first measurement is taken at week 4. All of these patients are in the active treatment arm, except for #27. On the other hand, the two patients influential for the measurement model parameters stay in the study up to week 24 and have no observation for the last measurement occasion at 1 year. Patient # 68 received the active treatment, and his/her visual acuity decreases substantially after week 4, thereafter staying more or

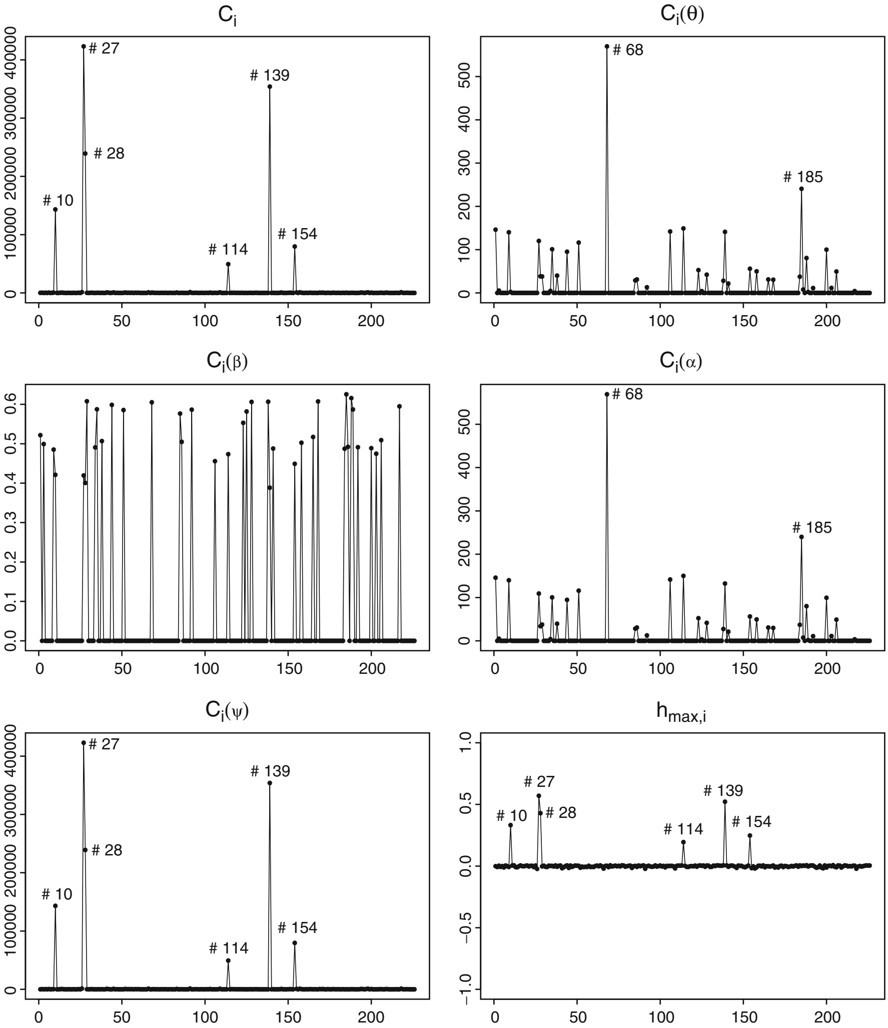


Fig. 34.7 Age-related macular degeneration trial. Index plots of C_i, $C_i(\vec{\theta})$, $C_i(\vec{\alpha})$, $C_i(\vec{\beta})$, $C_i(\vec{\psi})$ and of the components of the direction $\vec{h}_{\max,i}$ of maximal curvature

less level. Opposite, patient #185 is enrolled in the placebo treatment arm and his/her visual acuity increases after week 4, then sloping down a little after week 12.

It is interesting to consider an analysis without these influential observations. Therefore, we applied the selection model to three subsets of the data. The first subset was obtained by removing all the eight influential patients mentioned before. In the second subset of the data, patients #10, #27, #28, #114, #139, and #154 were removed, since these are overall the most influential ones. Finally, patients #68 and #185, who seemed to be influencing the measurement model the most, were

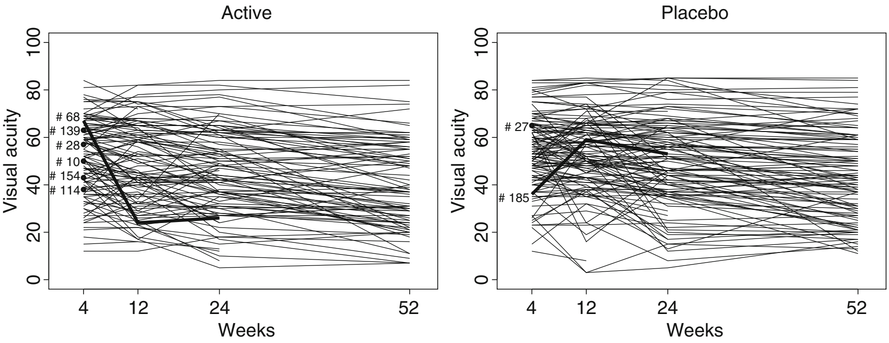


Fig. 34.8 Age-related macular degeneration trial. Individual profiles for both treatment arms, with influential subjects highlighted

removed, resulting in the third subset. The results of these analyses are shown in Table 34.10. We compare the results of the MAR and MNAR analyses.

After removing all influential patients (Set 1), the estimates of the dropout model parameters ψ_1 and ψ_2 are approximately the same, whereas the estimate of ψ_0 decreases from -1.86 to -1.90 under MAR, and from -1.81 to -1.85 under MNAR. The same can be seen after removing patients #10, #27, #28, # 114, #139, and #154, who have large overall C_i and $C(\vec{\psi})$ values (Set 2). Considering the treatment effect at 1 year, its estimate decreases from -5.18 to -5.45 under the MAR assumption, and from -5.71 to -6.12 under the MNAR assumption, resulting in a decrease of the p-value from 0.048 to 0.040 and from 0.030 to 0.021 under MAR and MNAR, respectively.

There is no impact on the likelihood ratio test for MAR against MNAR after removing all influential patients, the deviance G^2 only changes slightly from 2.5 to 2.6. If this likelihood ratio test would follow a standard χ^2_1-distribution, we would fail to reject the null hypothesis, which leads us to the MAR assumption. However, the test of MAR against MNAR is non-standard and it cannot be used as such (Rotnitzky et al. 2000; Jansen et al. 2006). Moreover, recall that one can never test for the assumption of MNAR versus MAR missingness.

Further, after removing the second set of influential patients, that is, patients #10, #27, #28, #114, #139, the estimate of the treatment effect at 1 year increases from -5.18 to -5.06 under the MAR analysis, yielding a slightly increased borderline p-value, whereas it decreases with 0.01 under the MNAR analysis and together with a decreased standard error the latter yields a small decrease in the p-value. The deviance G^2 for the likelihood ratio test for MAR against MNAR remains 2.5.

Finally, we perform the same analyses on the third set, with patients #68 and #185 removed. Both for the MAR and MNAR analysis, again the estimate of the treatment effect at 1 year decreases quite a lot, from -5.18 to -5.56 and from -5.71 to -6.09, respectively. Consequently, the p-value also drops down from 0.048 to 0.029 under

Table 34.10 Age-related macular degeneration trial. Parameter estimates (standard errors) explicitly modeling the missing data mechanism under MAR and MNAR assumptions, after removing the following subsets of subjects 10, 27, 28, 114, 139, 154, 68, 185 (Set 1); 10, 27, 28, 114, 139, 154 (Set 2); and 68, 185 (Set 3)

Subjects removed		Set 1				Set 2				Set 3			
		MAR		MNAR		MAR		MNAR		MAR		MNAR	
Effect	Parameter	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)
Measurement model													
Intercept 4	β_{11}	54.14	(1.51)	54.15	(1.49)	54.30	(1.47)	54.30	(1.46)	53.84	(1.48)	53.84	(1.47)
Intercept 12	β_{21}	53.09	(1.64)	53.06	(1.62)	53.16	(1.59)	53.13	(1.59)	52.94	(1.60)	52.91	(1.59)
Intercept 24	β_{31}	49.56	(1.77)	49.46	(1.75)	49.31	(1.74)	49.20	(1.72)	49.44	(1.73)	49.31	(1.72)
Intercept 52	β_{41}	44.40	(1.82)	43.97	(1.84)	44.00	(1.79)	43.58	(1.82)	44.38	(1.78)	43.90	(1.82)
Treatment 4	β_{12}	−3.13	(2.17)	−3.13	(2.11)	−3.28	(2.08)	−3.28	(2.06)	−2.95	(2.07)	−2.95	(2.05)
Treatment 12	β_{22}	−4.48	(2.36)	−4.63	(2.29)	−4.55	(2.26)	−4.69	(2.24)	−4.47	(2.26)	−4.60	(2.23)
Treatment 24	β_{32}	−3.80	(2.56)	−4.04	(2.49)	−3.55	(2.48)	−3.79	(2.44)	−3.85	(2.44)	−4.04	(2.42)
Treatment 52	β_{42}	−5.45	(2.66)	−6.12	(2.66)	−5.06	(2.59)	−5.72	(2.61)	−5.56	(2.55)	−6.09	(2.58)
Dropout model													
Intercept	ψ_0	−1.90	(0.47)	−1.85	(0.49)	−1.90	(0.47)	−1.85	(0.49)	−1.85	(0.46)	−1.81	(0.47)
Previous	ψ_1	−0.019	(0.010)	0.018	(0.022)	−0.019	(0.010)	0.017	(0.022)	−0.020	(0.009)	0.017	(0.022)
Current	ψ_2			−0.044	(0.024)			−0.043	(0.024)			−0.043	(0.024)
−2 log-likelihood		6,535.3		6,532.7		6,606.9		6,604.4		6,706.4		6,703.8	
Treatment at 1 year	p-value	0.040		0.021		0.051		0.028		0.029		0.018	

MAR and from 0.030 to 0.018 under the MNAR analysis. The deviance for the likelihood ratio test for MAR changes again from 2.5 to 2.6.

34.8 Conclusions

In this chapter, it has been shown that the use of simple methods, such as complete case analysis or last observation carried forward, at least require the missingness mechanism to be MCAR. Therefore, while historically very popular, these methods carry major drawbacks and can and ought to be replaced with more advanced methods. One such method is a likelihood-based ignorable analysis. It has a broad basis in the sense that it is valid under MAR and compatible with general strategies such as linear mixed models for Gaussian outcomes or generalized linear mixed models, within the random-effects model family, for non-Gaussian outcomes. These methods are simple to conduct as it would be in contexts where data are complete.

In the non-Gaussian setting, one often opts for a semi-parametric approach such as generalized estimating equations within the marginal model family. Since this method also requires the missingness to be MCAR, a reasonably straightforward modification is needed in order to make the method suitable for the MAR setting. Two alternatives have been considered in this chapter, that is, weighted generalized estimating equations and multiple-imputation-based generalized estimating equations. Both methods only require a little amount of programming which can be done using standard statistical software.

Thus, it has been made clear that a primary analysis should consist of methods which assume the missing data to be MAR instead of MCAR. On the other hand though, one can hardly ever rule out the possibility of missingness to be MNAR, which implies that the need may exist to consider MNAR models. Therefore, we have considered a full selection model proposed by Diggle and Kenward (1994) for Gaussian outcomes, as well as pattern-mixture models based on two strategies. Using the identifying restrictions strategy, a link exists between pattern-mixture models and selection models. Indeed, the ACMV identifying restriction is the pattern-mixture model equivalent of the MAR missingness assumption within the selection model framework. In addition, a comparison between the selection and pattern-mixture modeling approaches is useful to obtain additional insight into the data and/or to assess sensitivity.

Since MNAR models are based on assumptions regarding the unobserved outcomes, which are not verifiable from the available, observed data, one can never test the assumption of an MNAR model for or against MAR. This underscores that the conclusions drawn based on MNAR model are sensitive to the posited and unverifiable model assumptions. Consequently, in any incomplete data setting, there cannot be anything that could be called a definitive analysis and one should perform a sensitivity analysis. More confidence in the results can be gained if both models lead to similar conclusions. As seen in this chapter, this can be done not only within both the selection model and pattern-mixture model framework, but also across both frameworks. Further, we have discussed local influence, which is a sensitivity tool

Table 34.11 Age-related macular degeneration trial. Estimates, standard errors, and p-values for the treatment effect at 1 year for the ignorable direct-likelihood analysis, both Diggle–Kenward selection models assuming either MAR or MNAR, as well as the pattern-mixture model using ACMV restrictions

	Treatment at 1 year		
Model	Estimate	s.e.	p-value
Ignorable direct-likelihood	−5.18	2.59	0.047
Pattern-mixture model (ACMV)	−5.19	2.81	0.065
MAR selection model	−5.18	2.62	0.048
MNAR selection model	−5.71	2.63	0.030

that can be used within the selection model framework to detect (groups of) subjects which highly influence the obtained results.

All methods for analyzing incomplete data which are considered in this chapter are applied to the ARMD trial. In this way a broad sensitivity analysis is performed on this case study. For instance, let us focus on the treatment effect at 1 year obtained by the different methods, to assess the sensitivity of the modeling assumptions on the conclusions. A comparison of the estimates, standard errors, and p-values is provided in Table 34.11. Whereas the pattern-mixture model based on ACMV identifying restrictions shows a borderline (in)significant treatment effect at 1 year, it moves to borderline significant for the ignorable direct-likelihood analysis, which assumes MAR missingness, and the MAR Diggle–Kenward selection model, and its significance becomes more prominent in turning to the MNAR Diggle–Kenward selection model. In conclusion, the primary analysis, which ideally would be based on a model assuming MAR non-response, for instance, an ignorable direct-likelihood analysis as in this case, yields a borderline significant treatment effect at 1 year. To assess the sensitivity thereof, this primary analysis is extended with other MAR models under different modeling frameworks which clearly confirms this borderline significance. Expanding this sensitivity analysis by turning attention to MNAR models shows a decreased p-value thereby concluding a significant treatment effect at 1 year. Thus, a cautious conclusion is that there is some evidence for a treatment effect at the end of the study.

Acknowledgements We gratefully acknowledge support from FWO-Vlaanderen Research Project G.0002.98: “Sensitivity Analysis for Incomplete and Coarse Data” and from Belgian IUAP/PAI network “Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data”.

References

Aerts M, Geys H, Molenberghs G, Ryan LM (2002) Topics in modelling of clustered binary data. Chapman & Hall, London

Afifi A, Elashoff R (1966) Missing observations in multivariate statistics I: review of the literature. J Am Stat Assoc 61:595–604

Baker SG, Rosenberger WF, DerSimonian R (1992) Closed-form estimates for missing counts in two-way contingency tables. Stat Med 11:643–657

Beckman RJ, Nachtsheim CJ, Cook RD (1987) Diagnostics for mixed-model analysis of variance. Technometrics 29:413–426

Beunckens C, Sotto C, Molenberghs G (2008) A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. Comput Stat Data Anal 52:1533–1548

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88:9–25

Chatterjee S, Hadi AS (1988) Sensitivity analysis in linear regression. Wiley, New York

Cook RD (1977) Detection of influential observations in linear regression. Technometrics 19: 15–18

Cook RD (1979) Influential observations in linear regression. J Am Stat Assoc 74:169–174

Cook RD (1986) Assessment of local influence. J R Stat Soc Ser B 48:133–169

Cook RD, Weisberg S (1982) Residuals and influence in regression. Chapman & Hall, London

Dempster AP, Rubin DB (1983) Overview. In: Madow WG, Olkin I, Rubin DB (eds) Incomplete data in sample surveys. Theory and annotated bibliography, vol II. Academic, New York, pp 3–10

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J R Stat Soc Ser B 39:1–38

Diggle PJ, Kenward MG (1994) Informative drop-out in longitudinal data analysis (with discussion). Appl Stat 43:49–93

Diggle PJ, Heagerty P, Liang K-Y, Zeger SL (2002) Analysis of longitudinal data. Oxford University Press, New York

Fitzmaurice GM, Molenberghs G, Lipsitz SR (1995) Regression models for longitudinal binary responses with informative dropouts. J R Stat Soc Ser B 57:691–704

Glynn RJ, Laird NM, Rubin DB (1986) Selection modeling versus mixture modeling with nonignorable nonresponse. In: Wainer H (ed) Drawing inferences from self-selected samples. Springer, New York, pp 115–142

Hartley HO, Hocking R (1971) The analysis of incomplete data. Biometrics 27:7783–808

Heckman JJ (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Ann Econ Soc Meas 5:475–492

Hogan JW, Laird NM (1997) Mixture models for the joint distribution of repeated measures and event times. Stat Med 16:239–258

Ibrahim JG, Molenberghs G (2009) Missing data methods in longitudinal studies: a review (with discussion and rejoinder). Test 18, 68–80

Jansen I, Molenberghs G (2008) A flexible marginal modeling strategy for non-monotone missing data. J R Stat Soc Ser A 171:347–373

Jansen I, Hens N, Molenberghs G, Aerts M, Verbeke G, Kenward MG (2006) The nature of sensitivity in monotone missing not at random models. Comput Stat Data Anal 50:830–858

Kenward MG, Molenberghs G (1998) Likelihood based frequentist inference when data are missing at random. Stat Sci 12:236–247

Kenward MG, Molenberghs G (2009) Last observation carried forward: a crystal ball? J Biopharm Stat 19(5):872–888

Kenward MG, Molenberghs G, Thijs H (2003) Pattern-mixture models with proper time dependence. Biometrika 90:53–71

Laird NM (1994) Discussion to Diggle PJ, Kenward MG: informative dropout in longitudinal data analysis. Appl Stat 43:84

Laird NM, Ware JH (1982) Random effects models for longitudinal data. Biometrics 38:963–974

Lesaffre E, Verbeke G (1998) Local influence in linear mixed models. Biometrics 54:570–582

Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22

Little RJA (1993) Pattern-mixture models for multivariate incomplete data. J Am Stat Assoc 88:125–134

Little RJA (1994) A class of pattern-mixture models for normal incomplete data. Biometrika 81:471–483

Little RJA (1995) Modeling the drop-out mechanism in repeated measures studies. J Am Stat Assoc 90:1112–1121

Little RJA, Rubin DB (1987) Statistical analysis with missing data. Wiley, New York

Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York

Mallinckrodt CH, Clark WS, David SR (2001a) Type I error rates from mixed-effects model repeated measures versus fixed effects analysis of variance with missing values imputed via last observation carried forward. Drug Inform J 35:1215–1225

Mallinckrodt CH, Clark WS, David SR (2001b) Accounting for dropout bias using mixed-effects models. J Biopharm Stat Ser 11(1 & 2):9–21

Mallinckrodt CH, Clark WS, Carroll RJ, Molenberghs G (2003a) Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. J Biopharm Stat 13:179–190

Mallinckrodt CH, Sanger TM, Dube S, Debrota DJ, Molenberghs G, Carroll RJ, Zeigler Potter WM, Tollefson, GD (2003b) Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. Biol Psychiatry Ser 53:754–760

Meng XL (1994) Multiple-imputation inferences with uncongenial sources of input (with discussion). Stat Sci 9:538–573

Michiels B, Molenberghs G, Lipsitz SR (1999) A pattern-mixture odds ratio model for incomplete categorical data. Commun Stat Theory Methods 28:2843–2869

Michiels B, Molenberghs G, Bijnens L, Vangeneugden T, Thijs H (2002) Selection models and pattern-mixture models to analyze longitudinal quality of life data subject to dropout. Stat Med 21:1023–1041

Molenberghs G, Kenward MG (2007) Missing data in clinical studies. Wiley, New York

Molenberghs G, Verbeke G (2005) Models for discrete longitudinal data. Wiley, New York

Molenberghs G, Kenward MG, Lesaffre E (1997) The analysis of longitudinal ordinal data with non-random dropout. Biometrika 84:33–44

Molenberghs G, Michiels B, Kenward MG, Diggle PJ (1998) Missing data mechanisms and pattern-mixture models. Stat Neerl 52:153–161

Molenberghs G, Michiels B, Lipsitz SR (1999) A pattern-mixture odds ratio model for incomplete categorical data. Commun Stat Theory Methods 28:2843–2869

Molenberghs G, Verbeke G, Thijs H, Lesaffre E, Kenward MG (2001) Mastitis in dairy cattle: local influence to assess sensitivity of the dropout process. Comput Stat Data Anal 37:93–113

Murray GD, Findlay JG (1988) Correcting for the bias caused by drop-outs in hypertension trials. Stat Med 7:941–946

Nelder JA, Mead R (1965) A simplex method for function minimisation. Comput J 7:303–313

Neuhaus JM (1992) Statistical methods for longitudinal and clustered designs with binary responses. Stat Methods Med Res 1:249–273

Neuhaus JM, Kalbfleisch JD, Hauck WW (1991) A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. Int Stat Rev 59:25–35

Pharmacological Therapy for Macular Degeneration Study Group (1997) Interferon α-IIA is ineffective for patients with choroiadal neovascularization secondary to age-related macular degeneration. Arch Ophthalmol 115:865–872

Robins JM, Rotnitzky A, Zhao LP (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc 90:106–121

Robins JM, Rotnitzky A, Scharfstein DO (1998) Semiparametric regression for repeated outcomes with non-ignorable non-response. J Am Stat Assoc 93:1321–1339

Rotnitzky A, Cox DR, Bottai M, Robins J (2000) Likelihood-based inference with singular information matrix. Bernouilli 6:243–284

Rubin DB (1976) Inference and missing data. Biometrika 63:581–592

Rubin DB (1978) Multiple imputation in sample surveys – a phenomenological Bayesian approach to nonresponse. In: Imputation and editing of faulty or missing survey data. U.S. Department of Commerce, Washington, DC, pp 1–23

Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York

Rubin DB (1994) Discussion to Diggle PJ, Kenward MG: informative dropout in longitudinal data analysis. Appl Stat 43:80–82

Schafer JL (1997) Analysis of incomplete multivariate data. Chapman & Hall, London

Schafer JL (1999) Multiple imputation: a primer. Stat Methods Med Res 8:3–15

Sheiner LB, Beal SL, Dunne A (1997) Analysis of nonrandomly censored ordered categorical longitudinal data from analgesic trials. J Am Stat Assoc 92:1235–1244

Siddiqui O, Ali MW (1998) A comparison of the random-effects pattern-mixture model with last-observation-carried-forward (LOCF) analysis in longitudinal clinical trials with dropouts. J Biopharm Stat 8:545–563

Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation. J Am Stat Assoc 82:528–550

TenHave TR, Kunselman AR, Pulkstenis EP, Landis JR (1998) Mixed effects logistic regression models for longitudinal binary response data with informative dropout. Biometrics 54:367–383

Thijs H, Molenberghs G, Michiels B, Verbeke G, Curran D (2002) Strategies to fit pattern-mixture models. Biostatistics 3:245–265

Verbeke G, Molenberghs G (1997) Linear mixed models in practice: a SAS-oriented approach. Lecture notes in statistics 126. Springer, New York

Verbeke G, Molenberghs G (2000) Linear mixed models for longitudinal data. Springer, New York

Verbeke G, Lesaffre E, Spiessens B (2001a) The practical use of different strategies to handle dropout in longitudinal studies. Drug Inform J 35:419–439

Verbeke G, Molenberghs G, Thijs H, Lesaffre E, Kenward MG (2001b) Sensitivity analysis for non-random dropout: a local influence approach. Biometrics 57:7–14

Wu MC, Bailey KR (1988) Analysing changes in the presence of informative right censoring caused by death and withdrawal. Stat Med 7:337–346

Wu MC, Bailey KR (1989) Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. Biometrics 45:939–955

Wu MC, Carroll RJ (1988) Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics 44:175–188

Generalized Estimating Equations 35

Andreas Ziegler and Maren Vens

Contents

A. Ziegler
Institute of Medical Biometry and Statistics, University Medical Center Schleswig-Holstein, Campus Lübeck and Center for Clinical Trials, University of Lübeck, Lübeck, Germany

M. Vens
Institute of Medical Biometry and Statistics, University of Lübeck, University Medical Center Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_45,

35.1 Introduction

Generalized linear models (GLMs) are a standard regression approach for analyzing univariate non-normal data. In their breakthrough paper, Nelder and Wedderburn (1972) have derived GLM as a unifying approach for fitting models with dependent variables that are count data or dichotomous. GLM is nicely summarized in chapter ▸Regression Methods for Epidemiological Analysis of this handbook, the great introductory text book of Dobson (2001) or the excellent monograph by McCullagh and Nelder (1989). Here, the user specifies a link function to relate the independent and the dependent variables. For example, in epidemiology, the standard choice for dichotomous dependent variables is the logit link function to model the mean structure, a model which is known for more than 50 years. Another advantage of the GLM approach in this situation is that the variance function does not need to be explicitly specified. It is automatically generated from the assumption that binary data are assumed to be Bernoulli distributed.

In many applications, however, observations are not independent, for example, when measurements are taken repeatedly from the same subject over time or when paired data are taken, for example, from the left and the right eye. Neglecting dependencies in such a situation can lead to false conclusions. Specifically, the precision of the results and thereby their significance is usually overestimated. If one aims at using standard estimation techniques like maximum likelihood (ML) or least squares in such correlated data situations, one needs to specify the complete multivariate distribution. This seems to be simple for paired data, where one has to additionally specify a single correlation coefficient in addition to the mean and the variance. However, if three observations are taken from the same subject over time, not only three pair-wise correlations need to be specified in addition to the link function and the variance function but also one third-order moment. And this complexity increases exponentially with the number of observations. The unfavorable characteristic of ML and other standard methods in such situations is that their desirable statistical properties like asymptotic normality of the estimators rely on the correct specification of the complete multivariate distribution.

An interesting approach that does not require correct specification of the entire distribution is the generalized estimating equations (GEE). The key feature of GEE is that only the mean structure is crucial, in that it needs to be correctly specified. Valid inference about regression coefficients is possible even if the variances and correlations are misspecified. Variances and correlations are not of primary interest and therefore considered nuisance. Of course, if the user is also interested in the associations between observations from the same cluster, he/she needs to correctly specify the means and the variances and correlations. However, all higher-order moments are of minor interest even in this situation. As there is no free lunch, this robustness property of GEE on the one hand comes at the cost of a loss of efficiency. On the other hand, the robustness property makes the GEE approach very interesting and relatively simple for applications. The GEE has been further developed in many different ways since its introduction in 1986 by Liang and Zeger (Liang and Zeger 1986;

Zeger and Liang 1986), and it is these extensions which make the approach very flexible.

The outline of this contribution is as follows. We describe several scenarios which are typical for GEE analysis and demonstrate the simplicity of GEE in an introductory example. We discuss the interpretability of GEE models. This is followed by an overview of standard GEE1 models, that is, GEE for the mean structure and the discussion of different working correlation structures that are available in standard software packages. We illustrate how sample size and power calculations may be performed using available software. Finally, we illustrate the use of GEE in a real data example. Technical details are banned to two appendices. Specifically, the GEE are derived as a combination of generalized least squares (GLS) and GLM in Appendix 2.

35.2 Areas of Application

A series of applications using GEE were published in the last 2 decades. The following introductory examples aim at giving a flavor of the different areas of application and scientific problems that were studied.

35.2.1 Cohort Study

In a longitudinal cohort study, human papilloma virus (HPV) detection was investigated in a group of pregnant women and a group of age-frequency matched non-pregnant women. The aim of the study was to clarify previous conflicting data concerning the detection of HPV infection in the pregnant versus the non-pregnant state while adjusting for other known risk factors of cervical cancer. To this end, 5 follow-ups at 3-month intervals were conducted. Here, the correlation between observations at different time points was considered nuisance (Ziegler et al. 2003).

35.2.2 Household Data

This example is unusual for aggregation within households. In general, correlation between household members living in the same household needs to be taken into account. This standard example will be considered in detail in the next section. Here, we consider a randomized trial, in which the functionality of two different types of smoke alarms with two different types of batteries was investigated in rural households (Yang et al. 2008). Specifically, alarm function and false alarms were tested 12 months after installation. Because several alarms were installed in a single household, correlation within households had to be taken into account. Specifically, correct alarms within a household are correlated. Furthermore, false alarms might be correlated if batteries have similar charge levels.

35.2.3 Clinical Epidemiology: Parallel Group Design with Repeated Measurements

In a double-blind placebo-controlled randomized multicenter trial, the edema-protective effect of a vasoactive drug was investigated in patients suffering from chronic venous insufficiency after decongestion of the legs (See Sect. 35.8). Two-hundred-thirty-one patients were randomly assigned to medical compression stockings plus SB-LOT (90 mg coumarin and 540 mg troxerutin per day) or medical compression stockings plus placebo for the first 4 weeks and SB-LOT or placebo for the second 12 weeks of the study. The primary efficacy endpoint was the lower leg volume measured. In total, patients were followed-up 5 times: 4, 6, 8, 12, and 16 weeks after starting the drug therapy. The primary analysis was a baseline-adjusted (visit 0) covariance analysis (ANCOVA) of the difference of the leg volume at the final visit minus the volume at baseline, that is, the time point of randomization for demonstrating a difference of the vasoactive drug when compared with placebo (Vanscheidt et al. 2002). A secondary analysis which aimed at detecting a difference in the slopes using GEE will be presented in Sect. 35.8.

35.2.4 Clinical Epidemiology: Crossover Study

In 2 × 2 crossover trials, patients receive both the active drug and the placebo, but the order is randomized (cf. chapter ▶Pharmacoepidemiology of this handbook). If there is no period effect, the standard analysis is the McNemar test investigating whether more subjects have given a preference to the active treatment. One alternative for analysis is the GEE approach in which treatment responses from both time points are analyzed jointly while at the same time taking into account the correlatedness of the responses (Diggle et al. 1994). 2 × 2 crossover trials can therefore be considered as longitudinal studies with two time points. The standard example for a 2×2 crossover study with binary endpoint has been provided by Jones and Kenward (2003). It consisted in safety data from a placebo-controlled bi-center trial on cerebrovascular deficiency. The primary endpoint of the study was whether the electrocardiogram as judged by a cardiologist was considered to be normal or abnormal. It will be used as example several times in this chapter and for the first time in Example 35.2.

35.2.5 Clinical Epidemiology: Laterality in Ophthalmological Studies

In the Blue Mountains Eye Study, a population-based cohort study, participants were investigated at baseline, after 5 and 10 years to determine whether local nutritional or ischemic factors are involved in cataract pathogenesis

(Tan et al. 2008). Specifically, the investigators assessed whether narrowed retinal vessel caliber predicted the long-term incidence of age-related cataract. Measurements were taken for both eyes so that two levels of nesting were present. The first level was caused by the time course and the second by the laterality, where both time course and laterality were considered nuisance.

35.2.6 Clinical Epidemiology: Local Correlation of Teeth in Dental Studies

In a placebo-controlled randomized clinical trial, the effect of a mouthrinse on gingivitis and tooth whitening was investigated (Hasturk et al. 2004). The first phase (1 month) was the experimental gingivitis phase and the second phase (5 months) was the oral hygiene phase, which included rinsing. Primary endpoints were chosen to reflect the gingival health and tooth whiteness. GEE were used to accommodate correlations between teeth from the same patient. The authors concluded that the fluoridated mouthrinse whitens teeth and reduces gingivitis.

35.2.7 Toxicological Litter Experiments

Ryan (1992) reanalyzed data from a developmental toxicity study in mice using GEE. The experiment included a control group and four groups exposed with different doses of diethylhexylphthalate (DEHP). Correlation between observations on the individual level arises from the litter. The authors concluded that the proportion of malformation in the offspring increased with dose. They also concluded that the intra-litter correlation was not constant.

35.2.8 Spatial Structure

In forest damage surveys, the state of tree, for example, the degree of defoliation, is measured in ordered categories. A typical survey uses a grid with rectangular meshes over a map in the survey area. For each grid point, damage states and covariates are measured for a fixed number of trees next to the grid point. A correct analysis needs to take into account the spatial correlation arising from neighboring effects among trees around the grid point (Fahrmeir and Pritscher 1996).

35.2.9 Diagnostic Tests

The evaluation of a new medical diagnostic test may focus on two different scientific questions.

Replacing a Test by a New One First, the new test may replace an existing one because of lower cost or higher validity. A related question would be the selection of the best test(s) from a bundle of new or established measurements. Of course, the tests are correlated in this setting, but the correlation is considered nuisance.

Using a New Test Supplementary to Existing Tests In the second situation, the new diagnostic test may be used supplementary to other new or established procedures. Here, only combinations of tests showing no or only a moderate correlation yield additional information. Subsequently, the association between tests is of primary interest in this setting (Martus et al. 2004).

35.2.10 Cluster of Individuals: Familial Aggregation

Genetic factors contribute to the etiology of many common diseases. While one goal of genetic epidemiological studies (cf. chapter ▶Statistical Methods in Genetic Epidemiology of this handbook) is to locate susceptibility genes for these complex diseases, it is important that strong evidence of familial aggregation be established at an early stage of research. GEE may be used for detecting familial aggregation and for investigating whether environmental or demographic factors influence these associations (Ziegler et al. 2000).

More generally, any cluster of individuals, including members of a household, children within a class at a school, or patients admitted to a general practitioner in a cluster randomized trial, yield correlated observations (cf. chapter ▶Cluster Randomized Trials of this handbook).

35.2.11 Summary

In all but the last two examples, the mean structure was of primary interest, while it was the association in the final two examples. Although the primary aim was the investigation of the mean structure in the first examples, the correlation within clusters should not be neglected for obtaining correct standard errors. With these various applications, we have shown that clustering between observations may arise from many different sources, including repeated measurements, households, or laterality. Cluster levels may be nested as demonstrated in the example from ophthalmology.

Correlations between sampling units, for example, the subjects in a household, the eyes of a subject, or the time point specific measurement, usually show a positive correlation (Cochran 1963). Examples of negative correlation are uncommon in epidemiology and biostatistics, but the examples given above encounter two. The first is related to litter experiments. Fetuses in a litter might be "competing" with respect to nutrition or other factors like space. The correlation typically depends on litter size, and it might be negative between fetuses according to fetus size. A similar example is the birthweight of human twins (Hanley et al. 2000). In the forest survey example, a spatial periodic variation can be observed (Baradat et al. 1996), and in some examples, it might be described by simple sine curves (Cochran 1963, pp. 218–219). The correlation between trees may therefore be negative. Simply speaking, a huge tree may show a negative correlation to the neighboring tree which might be small. The tree next to the small one may be big like the first tree, resulting in a positive correlation between the first and the third tree. In total, there might be a periodic correlation.

35.3 Introductory Example

In the previous section, we discussed many different areas of application for GEE. In this section, we demonstrate the effect of ignoring the correlation between sampling units. Using a simple example, we first assume that persons living in the same household are independent. Next, we consider a simple approach that allows to adjust for correlation by utilizing the sample size inflation factor (*SSIF*). This approach is, however, not entirely correct and cannot be easily generalized to more complex situations. Finally, we show that a simple GEE analysis yields very similar results as the *SSIF* approach. In contrast to the *SSIF* method, the GEE approach can be generalized easily and is very flexible.

Consider the raw data provided in Table 35.1. Clusters, that is, households, range in size from 1 to 7. Individuals in each of the 30 households were classified as to whether they had consulted a dentist in the past 12 months. A total of $N = 105$ subjects lived in the $n = 30$ households, giving an average household size of $T_a = 3.5$. Fifty-five persons (52.38%) consulted a dentist in the previous year. When we ignore the clustered nature of the data, we can easily estimate the standard error of the mean as usual by $\text{S.E.} = \hat{\sigma} \big/ \sqrt{N} = 0.0490$. The common asymptotic 95% confidence interval (CI) (chapter ▸Sample Size Determination in Epidemiological Studies of this handbook, Eq. 28.2) thus gives (42.78%–61.97%).

If subjects are correlated, the variability is altered. For example, when all subjects in a household either visit a dentist or do not visit a dentist, then the information from just one of the subjects living in the household is sufficient. By interviewing all household members, the sample size is just inflated but the additional subjects do not yield any additional information. If the correlation is not perfect but different from 0, interviewing more than one subject per household gives additional information but not as much as a subject from a new household. Still, there is some inflation in the sample size. If the correlation ϱ between all subjects in a household is identical

Table 35.1 Data on dentist visits for a cluster sample of 30 households

	Households																														Total
n_h	5	6	3	3	2	3	3	3	4	4	2	7	3	4	3	5	4	4	3	3	4	3	3	1	2	4	3	4	2	5	105
n_v	3	5	0	2	1	1	2	1	2	3	1	1	2	3	2	3	3	3	2	1	1	2	1	0	1	2	1	2	1	3	55

n_h number of subjects in household h, n_v number who had visited a dentist in the previous year (Modified from Cochran 1963; Hanley et al. 2000) and if the households are of equal size T, the *sample size inflation (SSIF)* can be shown to be (see Appendix 1)

$$SSIF = 1 + \varrho\,(T - 1)\,, \tag{35.1}$$

and the standard error of the mean correcting for the correlatedness of the data is S.E.$_{\cdot c} = \hat{\sigma}\sqrt{\widehat{SSIF}}\big/\sqrt{N}$ with $N = n\,T$. First, consider the extreme case $\varrho = 0$ of uncorrelated observations, then the *SSIF* equals 1, and no inflation occurs. If at the other extreme, $\varrho = 1$, the *SSIF* equals T, and no information is gained. If the cluster size varies, a simple approximation of T is through the average cluster size T_a.

In our example, the estimated Pearson correlation coefficient is $\hat{\varrho} = -0.07$, thus slightly negative. The *SSIF* is 0.825 so that the standard error corrected for correlation should be approximately S.E.$_{\cdot c} = \hat{\sigma}\sqrt{\widehat{SSIF}}\big/\sqrt{N} \approx 0.0446$, with (43.64%–61.14%) being the corresponding 95% CI. This CI is narrower than the CI ignoring the clustered nature of the data, which was 42.78%–61.97%.

Although the analysis using the *SSIF* is appealing and simple, it is not entirely correct because it relies on two assumptions. First, the correlations between all pairs of subjects living within a household need to be correctly specified. Here, we used the assumption that this correlation is identical. It may, however, vary between household members. For example, it is reasonable that children visit a dentist at least on an annual basis, while adults might not do so. Second, we used the average household size for calculating the SSIF. Household size was, however, not constant, and the degree of information between households of varying size may differ substantially.

In summary, there should be a simple robust approach, with robust meaning that it is valid irrespective of (a) the true correlation structure and (b) the cluster, that is, the household size. The independence estimating equation (IEE), a special and very simple type of GEE analysis, is able to accomplish both goals. If we employ the IEE, we assume that the subjects within a household are independent. As discussed before, this assumption does not hold, but by using a different variance estimator, the GEE approach "automatically" corrects for any misspecification of the variances and/or covariances. In the example, the IEE yield the same point estimate 52.38% for the proportion of subjects who visited a dentist in the previous year, the estimated standard error is 0.0436, and the corresponding 95% CI is given by (43.83%–60.93%). This CI is similar to the CI as obtained by the *SSIF* approach and about 2% narrower than the CI obtained by the naïve method.

IEE and GEE are implemented in various software packages, including SAS `PROC GENMOD`. The code for this simple example is

```
PROC GENMOD;
  CLASS household;
  MODEL visited = ;
  REPEATED SUBJECT = household /TYPE = ind;
run;
```

For application, data are arranged with one subject per row in a data matrix, `household` is a variable indicating the number of the household, and `visited` is a dichotomous variable being 1 if a subject visited a dentist and 0 otherwise. `TYPE = ind`, where `ind` is the abbreviation for "independent," is the option for calling the IEE.

35.4 Population-Averaged Interpretation

The GEE approach differs in fundamental conceptual ways from other models termed "state dependence" or "habit persistence" models. In state dependence models, also termed "conditional models," the response of an observation at time t may depend on responses within the same cluster, not only on some explanatory variables. In habit persistence models, also termed "binary choice models," which are commonly used in econometrics, a similar idea is followed. Here, however, latent variables are modeled which are interpreted as disposition or utility. In analogy to state dependence models, the utility at time point t may depend on the utility at other time points. Very popular in epidemiology are random effect models, where between-cluster variability is explicitly modeled and estimated. For a thorough discussion of these models, the reader may refer to the literature (see, for example, Fahrmeir and Tutz 1994; Lechner et al. 2008).

The GEE approach does not explicitly model between-cluster variation; instead, it focuses on the within-cluster similarity and then uses this estimated correlation to reestimate the regression parameters and to calculate standard errors. However, because the GEE approach does not explicitly model the between-cluster variation, the interpretation of the parameters is that of a population-averaged model, and this will be explained below using an example from Hanley et al. (2003) and Gail et al. (1984). To this end, consider a hypothetical study with extreme variation in the response probability from some clusters to others (Table 35.2). For simplicity, all clusters are of equal size, and exactly 50% of the subjects are exposed. Specifically, suppose that for half of the clusters the response proportion in unexposed subjects is 10%, while it is 50% in the exposed. For the other half of clusters, assume that 50% of the unexposed and 90% of the exposed subjects respond. The difference is 40 percentage points in both groups of clusters, and the odds ratio (*OR*) is 9.0. When the data are aggregated, the difference still is 40 percentage points, but the *OR* is as low as 5.4.

Table 35.2 Hypothetical data for illustrating the interpretation of generalized estimating equations

Clusters	Unexposed (response in %)	Exposed (response in %)	Difference (percentage points)	Odds ratio
1 to $n/2$	10	50	40	9.0
$n/2 + 1$ to n	50	90	40	9.0
Aggregated	30	70	40	5.4

The standard Cochran-Mantel-Haenszel type of analysis "recovers" the within-cluster *OR* of 9.0, as does a logistic regression, when an indicator variable is included for each cluster. Alternatively, one may fit a conditional logistic regression with cluster as stratum variable. However, the GEE approach yields an *OR* of 5.4 as for the aggregated data in Table 35.2. This *OR* of 5.4 contrasts the response probability for an exposed subject selected randomly from the *population* with the response probability for an unexposed subjected also selected randomly from the *population*. The random selection ignores the between-cluster variability and does not match on cluster. This *population-averaged* measure, from the marginal model used in the GEE approach, is specific to the mix of clusters studied. The *OR* of 9.0 contrasts the response probability for an exposed subject with the response probability for an unexposed subject *from the same cluster*, and it is therefore a *subject-specific model*. Both models have their own value, and the choice of the model depends on the interpretation required by the investigator.

35.5 Elements of a Generalized Estimating Equations (GEE) Analysis

The code piece shown at the end of Sect. 35.3 is a simple piece of code. In fact, the five lines already point to implicit assumptions that are made because some elements are not explicitly stated so that the default values are used. The user needs to explicitly or implicitly specify several components for fitting a GEE model. In detail, the following different elements (Box 35.1) need to be chosen by the researcher in any GEE software package (modified from Ballinger 2004; Stokes 1999):

Box 35.1. Important elements for a GEE analysis

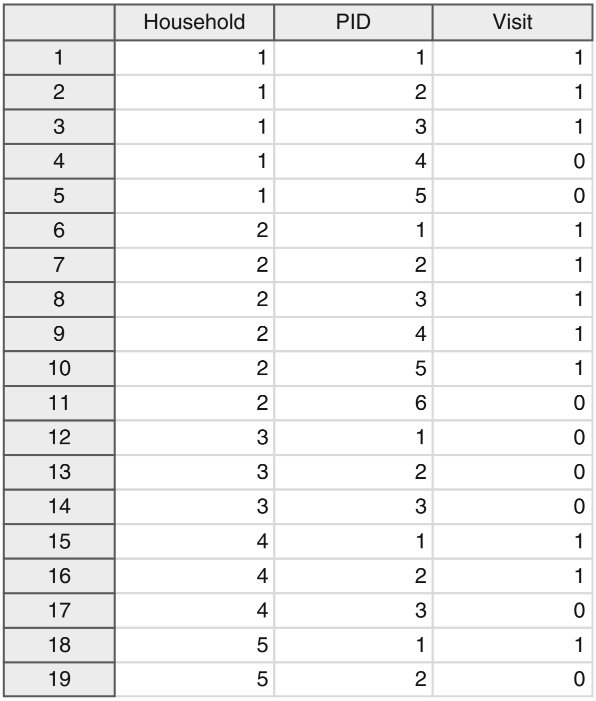
1. *Specify the dependent variable.* Mandatory. The dependent variable may be of a type that is suitable for estimation in a GLM. For example, it may be continuous, categorical (ordered or unordered), or dichotomous.
2. *Specify the independent variable(s).* Mandatory but may be left blank. This specification includes coding of variables and modeling of interaction terms.
3. *Specify the link function.* Optional; default: identity link. The link function connects the dependent and the independent variables. For dichotomous dependent variables, the logit link or the probit link are standard choices.
4. *Specify the clustering variable.* Mandatory. The clustering variable indicates the clustering of the dependent variable in the data. Examples for cluster levels include household, family, or subject in case of repeated measurements or laterality.
5. *Specify the distribution of the dependent variable.* Optional; default: normal distribution. Here, the standard option is generally followed; for

example, the binomial distribution is often used for dichotomous variables. However, several choices are possible as will be explained below. The choice of the distribution determines the variance function.

6. *Specify the correlation structure.* Mandatory. Observations within clusters are correlated. The efficiency of estimates may be improved by choosing a correlation structure that is reasonable for the user and that is close to the true correlation structure. Different options for the correlation structure and standard choices will be discussed below.
7. *Specify the statistical tests to be calculated.* Optional; default: no complex statistical tests. In some applications, users are interested in testing more complex hypotheses, and test statistics may be formulated, for example, by contrasts.
8. *Specify the confidence intervals to be estimated.* Optional; default: no complex confidence intervals. In some applications, users are interested in estimating more complex confidence intervals. These can be formulated in some GEE software packages.

Example 35.1. Elements of GEE analysis for introductory example
For the introductory example from Sect. 35.3, data were prepared in a table as shown in Fig. 35.1, and we made the following choices: The dependent variables was `visited`, no independent variables were chosen so that `PROC GENMOD` fitted a model with a regression constant. No link function was specified. Since the identity link $f(x) = x$ is the default, a mean was estimated. Note that we did not choose the logit link although the dependent variable is dichotomous because the aim was to estimate a proportion, that is, the mean of 0–1 variables. The clustering variable was the household. This is indicated by the `REPEATED SUBJECT` statement. The distribution of the dependent variable and the variance function were not specified. Default distribution is the normal distribution in SAS, and the default variance estimator is the standard empirical variance. The option `TYPE = ind` indicates that the independence correlation was chosen, that is, observations were assumed to be uncorrelated. The correction for the within-household correlation was automatically done in `PROC GENMOD` by choosing the variance estimator of the GEE.

Example 35.2. Crossover study
The data from this 2 × 2 crossover example have been utilized for illustration in a series of papers and books on GEE (see, e.g., Jones and Kenward 2003). A total of 67 patients from one center have been treated with both placebo (A, Treat = 0) and an active drug (B, Treat = 1). The order AB or BA in which patients received the two treatments was randomized, and the time point variable is coded as Time = 1 if period 2 and Time = 0 if period 1. The endpoint was whether the electrocardiogram (ECG) was considered to be normal (ECG = 1) or abnormal (ECG = 0). The subject indicating variable is PID. The summary data are displayed in Table 35.3. Thirty-four patients first received the active drug, followed by placebo. Twenty-eight patients had a normal ECG in the first period, only 22 in the second. Four patients with treatment order AB had an abnormal ECG under placebo but a normal ECG when treated with the active drug. The same response pattern, that is, normal ECG for B, abnormal for A, was observed for 6 patients with treatment order BA. Only two patients had a normal ECG under placebo and an abnormal ECG under the active drug, and these two patients had treatment order AB.



	Household	PID	Visit
1	1	1	1
2	1	2	1
3	1	3	1
4	1	4	0
5	1	5	0
6	2	1	1
7	2	2	1
8	2	3	1
9	2	4	1
10	2	5	1
11	2	6	0
12	3	1	0
13	3	2	0
14	3	3	0
15	4	1	1
16	4	2	1
17	4	3	0
18	5	1	1
19	5	2	0

Fig. 35.1 Snapshot of household data from SAS. Displayed is a part of the 30 families. The variable `Household` indicates the household with values from 1 to 30. `PID` is the person identifier within a household, is consecutively numbered from 1 for every household. `Visit` equals 1 if the subject has seen a dentist in the last year, 0 otherwise

Table 35.3 Data from 2 × 2 crossover study

	Responses					Period	
Group	1 1	0 1	1 0	0 0	Total	1	2
BA	22	0	6	6	34	28	22
AB	18	4	2	9	33	20	22

A placebo, B active treatment, BA patients who first received the active treatment, followed by placebo, AB patients who first received placebo, followed by the active treatment (Data from Jones and Kenward 1989)

Subsequently, the important elements of a GEE analysis are:

1 *Dependent variable*: `ecg`.
2 *Independent variables*: `treat` and `time`.
3 *Link function*: The logit link is the natural choice (`LINK = logit`).
4 *Clustering variable*: `pid`.
5 *Distribution of the dependent variable*: Binomial (`DIST = bin`).

6 *Correlation structure*: The standard choice is the independent working correlation (`TYPE = ind`).
7 *Statistical tests to be calculated*: No specific test contrasts required.
8 *Confidence intervals to be estimated*: No specific confidence intervals required.

The SAS code for this example with the choices from above is:

```
PROC GENMOD;
  CLASS pid;
  MODEL ecg = treat time /DIST = bin LINK = logit;
  REPEATED SUBJECT = pid /TYPE = ind;
run;
```

35.6 The Working Correlation Matrix

In this section, we consider common choices of the working correlation matrix, and we illustrate the use of different working correlation matrices with data from the simple 2 × 2 crossover study of Example 35.2. Let the correlation between subjects t and t' of cluster i be $\varrho_{itt'} = \mathbb{C}orr(y_{it}, y_{it'})$. The elements $\varrho_{itt'}$ are summarized to the so-called working correlation matrix. This working correlation matrix need not be correctly specified. It is just used for improving efficiency. Specifically, the closer the *working correlation* is to the *true correlation*, the more efficient the estimates are. Throughout, we assume that different clusters are uncorrelated so that $\mathbb{C}orr(y_{it}, y_{jt'}) = 0$ for $i \neq j$.

Standard choices for working correlation matrices in standard software packages are

- Fixed
- Independent
- Exchangeable
- Autoregressive
- Stationary
- Stationary m-dependent
- Unstructured
- m-dependent

Below, we define the different choices and also consider several working correlation structures that are not standard in all packages.

Fixed Working Correlation Structure A correlation structure which is simple but rarely used in applications is the *fixed* or *user defined working correlation* structure (common abbreviations: FIX, FIXED, USER). Here, the researcher has to pre-specify the entire structure and all values of the correlation matrix. Because correlations between observations are typically unknown a priori in practice, the fixed working correlation structure is rarely used.

Independent Working Correlation Structure The simplest structure is the *independent working correlation* structure (common abbreviations: IND, INDE, INDEP). Here,

$$\mathbb{C}orr(y_{it}, y_{it'}) = \begin{cases} 1, & \text{if } t = t' \\ 0, & \text{if } t \neq t' \end{cases} .$$

No correlation parameter needs to be specified in this case, and the example correlation matrix is

$$\begin{pmatrix} 1 & & & \\ 0 & 1 & & \\ \vdots & \ddots & \ddots & \\ 0 & \cdots & 0 & 1 \end{pmatrix} .$$

This correlation structure reflects the situation in which one assumes that there is no correlation between responses in the same cluster.

Exchangeable Working Correlation Structure The *exchangeable working correlation* structure, also termed *compound symmetry working correlation* structure, is a natural choice in family studies, household studies, that is, in the case of cluster sampling (common abbreviations: EX, EXCH, CS). Here, one assumes that all correlations within a cluster are equal, that is,

$$\mathbb{C}orr(y_{it}, y_{it'}) = \begin{cases} 1, & \text{if } t = t' \\ \varrho, & \text{if } t \neq t' \end{cases} .$$

The correlation matrix has the form

$$\begin{pmatrix} 1 & & & \\ \varrho & 1 & & \\ \vdots & \ddots & \ddots & \\ \varrho & \cdots & \varrho & 1 \end{pmatrix} .$$

The number of parameters to be estimated is 1.

Although this working correlation structure assumes that all observations within a cluster have the same correlation, it is often a good choice even if the true correlation is not identical for all observations in a cluster and between clusters.

Autoregressive Working Correlation Structure A more reasonable working correlation structure than the m-dependent working correlation is the *autoregressive working correlation* (common abbreviations: AR, AR(1)). Specifically, autoregressive refers to the autoregressive model of order 1. The corresponding working correlation structure is given by

$$\mathbb{C}orr(y_{it}, y_{it'}) = \begin{cases} 1, & \text{if } t = t' \\ \varrho^{|t-t'|}, & \text{if } t \neq t' \end{cases} .$$

The correlation matrix thus has the form

$$\begin{pmatrix} 1 & & & \\ \varrho & 1 & & \\ \vdots & \ddots & \ddots & \\ \varrho^{T-1} & \cdots & \varrho & 1 \end{pmatrix} .$$

The number of parameters to be estimated is 1. This working correlation matrix reflects that all observations are correlated. However, there is an exponential decay of the correlation over time.

Stationary Working Correlation Structure Another working correlation structure that is of interest for use with longitudinal data is the *stationary working correlation* (common abbreviations: STA, STAT). Here, it is assumed that all measurements with a specific distance in time have equal correlations. Specifically, it is assumed that observations 1 and 3 have the same correlation as observations 2 and 4 or 5 and 7, and the general definition of the stationary working correlation is

$$\mathbb{C}orr(y_{it}, y_{it'}) = \begin{cases} 1, & \text{if } t = t' \\ \varrho_{|t-t'|}, & \text{if } t \neq t' \end{cases} .$$

The correlation matrix thus has the form

$$\begin{pmatrix} 1 & & & \\ \varrho & 1 & & \\ \vdots & \ddots & \ddots & \\ \varrho_{T-1} & \cdots & \varrho & 1 \end{pmatrix} ,$$

and the number of parameters to be estimated is $T - 1$.

Stationary *m*-Dependent Working Correlation Structure The general definition of the *m-dependent stationary working correlation* structure is

$$\mathbb{C}orr(y_{it}, y_{it'}) = \begin{cases} 1, & \text{if } t = t' \\ \varrho_{|t-t'|}, & \text{if } t \neq t' \text{ and } |t-t'| \leq m \\ 0, & \text{if } |t-t'| > m \end{cases} .$$

The most commonly used form is the 1-dependent working correlation, where adjacent observations are correlated with ϱ_1. All other observations are assumed to be uncorrelated. Its common abbreviation is MDEP(m), where m is a number for the depth. Unfortunately, this working correlation is called m-dependent working correlation structure in most software packages despite its stationarity.

Unstructured Working Correlation The final common working correlation does not make any assumption on a specific structure, and it is therefore called *unstructured working correlation* (common abbreviations: UN, UNSTR). It is defined as

$$\mathbb{C}orr(y_{it}, y_{it'}) = \begin{cases} 1, & \text{if } t = t' \\ \varrho_{tt'}, & \text{if } t \neq t' \end{cases},$$

and the correlation matrix has the form

$$\begin{pmatrix} 1 & & & \\ \varrho_{1,2} & 1 & & \\ \vdots & \ddots & \ddots & \\ \varrho_{1,T} & \cdots & \varrho_{T-1,T} & 1 \end{pmatrix}.$$

The number of parameters to be estimated is $T(T-1)/2$.

This correlation structure involves a large number of parameters that need to be estimated, and it is therefore useful only if there is a natural ordering of the observations within a cluster, for example, because of a longitudinal setting. Furthermore, cluster sizes should be similar because it is not reasonable to estimate a correlation coefficient from one or two pairs of observations. One ideal setting would be a longitudinal study with an equal number of follow-ups such as in the SB-LOT study of Sect. 35.2.3. Here, a natural ordering is given by the time points, and for all patients the same number of follow-ups was considered.

***m*-Dependent Working Correlation Structure** The *m-dependent working correlation* structure is not often used in applications. The assumption is that there is a band of different correlations. All correlations are truncated to zero after the m^{th} band. All other correlations equal the unstructured working correlation. The definition of the m-dependent correlation is

$$\mathbb{C}orr(y_{it}, y_{it'}) = \begin{cases} 1, & \text{if } t = t' \\ \varrho_{tt'}, & \text{if } t < t' \text{ and } |t - t'| \leq m \\ 0, & \text{if } |t - t'| > m \end{cases},$$

and the correlation matrix thus has the form

$$\begin{pmatrix} 1 & & & & & \\ \varrho_{1,2} & 1 & & & & \\ \vdots & \ddots & \ddots & & & \\ \varrho_{1,m} & & \ddots & 1 & & \\ & \ddots & & \ddots & \ddots & \\ \mathbf{0} & & \varrho_{T-m-1,T} & \cdots & \varrho_{T-1,T} & 1 \end{pmatrix} .$$

Nested Working Correlation Standard packages typically do not offer *nested correlation structures* as standard choice for a working correlation. In nested correlation structures, different levels of nesting are present. For example, in dental studies, teeth are nested within patients, and sites around a tooth are nested within teeth. A simple working correlation structure in this situation is

$$\mathbb{C}orr(y_{ie_1t}, y_{ie_2t'}) = \begin{cases} 1, & \text{if } t = t' \text{ and } e_1 = e_2 \\ \varrho_1, & \text{if } t = t' \text{ and } e_1 \neq e_2 \\ \varrho_2, & \text{if } t \neq t' \end{cases} .$$

Here, e_1, e_2 denote, for example, surfaces on the same tooth. Therefore, there might be a correlation ϱ_1 different from 0 between two surfaces e_1 and e_2 at the same tooth t. For $\boldsymbol{y}_i = (y_{i11}, y_{i21}, y_{i12}, y_{i22})'$, an example correlation matrix is given by

$$\begin{pmatrix} 1 & & & \\ \varrho_1 & 1 & & \\ \varrho_2 & \varrho_2 & 1 & \\ \varrho_2 & \varrho_2 & \varrho_1 & 1 \end{pmatrix} .$$

Nested correlation structures may also occur in repeated measurement studies. For example, repeated measurements are nested within a patient, and measurements of both eyes are nested within a measurement time point. As a consequence, there might thus be a correlation between the left and the right eye but also over time. Here, a combination of two different autoregressive structures seems to be appealing. Specifically,

$$\mathbb{C}orr(y_{ie_1t}, y_{ie_2t'}) = \begin{cases} 1, & \text{if } t = t' \text{ and } e_1 = e_2 \\ \varrho_1, & \text{if } t = t' \text{ and } e_1 \neq e_2 \\ \varrho_2^{|t-t'|}, & \text{if } t \neq t' \text{ and } e_1 = e_2 \\ \varrho_3^{|t-t'|}, & \text{if } t \neq t' \text{ and } e_1 \neq e_2 \end{cases}$$

might be an attractive choice in this situation because the correlation decreases over time. For measurements at the same eye, the correlation follows an AR(1) model and thus equals $\varrho_2^{|t-t'|}$. For different eyes at different time points, the correlation also decreases according to an AR(1) model but with a different correlation parameter $\varrho_3^{|t-t'|}$. A difficult choice in this example is the correlation for $t = t'$ and $e_1 \neq e_2$ which is different from 1 and therefore is not conform with the standard AR(1) model, where it should equal 1.

Example 35.3. Crossover study
Return to the data of the 2 × 2 crossover study from Example 35.2. To avoid the correlation between time points, we first analyze the data time point specific using a logistic regression model. Twenty-eight patients out of 34 receiving the active treatment had a normal ECG at the first time point, while only 20 out of 33 had a normal ECG at the first follow-up. At the second follow-up, 22 patients in both groups had a normal ECG. This time, 33 patients were in the active drug and 34 in the placebo group. For the time point specific analyses, the most reasonable choice is a logistic regression model which can be called in SAS using `PROC LOGISTIC`. The data have been named `jones` according to the first author of the book in which the example data have been provided (Jones and Kenward 2003). The model for the dependent variable `ecg` only includes the dummy-coded, that is, 0–1-coded treatment variable `treat`. Using a data step command like

```
if time = 2 then delete;
```

we have already deleted the second time point. The code for the logistic regression model is

```
PROC LOGISTIC DATA=jones;
MODEL ecg = treat;
run;
```

The log odds ratio (log *OR*) of the treatment effect is −1.1097, its standard error 0.5738, and the p-value 0.0531. The *OR* (95% CI in parenthesis) is 0.330 (0.107–1.015). Thus, there is a tendency for a superiority of the active treatment over placebo. But when only the first time point is considered, it is not significant at the 5% test level.
Similarly, when only the second time point is considered, the finding is not significant. Specifically, the *OR* (95% CI in parenthesis) is 0.917 (0.334–2.515), and the p-value is 0.8658.

With the GEE, the data from both time points can be combined and analyzed jointly. At the same time, the correlation between observations at time points 1 and 2 is adequately taken into account. This is not the case for the logistic regression. For the logistic regression model

```
PROC LOGISTIC DATA=jones;
MODEL ecg = time treat;
run;
```

which considers both the time point and the treatment effect but ignores the correlation of observations over time, the *OR* (95% CI) for the treatment effect is 0.571 (0.273–1.202), and the p-value 0.1403. The correlation between responses and the time point is slightly negative ($\hat{\varrho} = -0.0643$) as in the introductory example of Sect. 35.3. We can therefore expect an decrease of the p-value when the correlation is adequately taken into account by using the independent working correlation matrix.
As already indicated above, the distribution chosen is the binomial distribution. This leads to the option `dist = b` in `PROC GENMOD`. The link function is the logit link, indicated in the syntax through the option `link = logit`. The cluster variable is `pid`. It needs

to be used twice in the SAS code. First, it has to be declared as a class variable `CLASS pid`. Second, it has to be declared in the `REPEATED` statement. Unfortunately, SAS does not use a "cluster statement" but a "repeated statement." This statement has to be used even if the data are of a different clustered nature than repeated. For example, in dental studies, measurements between different teeth are commonly correlated, and the repeated subject variable is the patient. In the present example, there is indeed a repetition of the ECG measurement over time.

In the simplest case, the working correlation is assumed to be the independent working correlation, and the corresponding option is `TYPE = ind`. The SAS code therefore is

```
PROC GENMOD DATA=jones;
CLASS pid;
MODEL ecg = time treat /dist = b link = logit;
REPEATED SUBJECT = pid /TYPE = ind;
run;
```

This analysis results in a significant treatment effect, and the *OR* (95% CI) is 0.572 (0.362–0.904) with $p = 0.0167$.

The results are similar when the probit link is used. In the code given above, only the option `link = logit` needs to be replaced by `link = probit`. Here, the parameter estimates cannot be interpreted in terms of log *OR*s or *OR*s. The p-value is, however, very similar with $p = 0.0149$.

Because the correlation coefficient between the responses and the time point, that is, over time, is approximately -0.06, we can also estimate the GEE with a fixed working correlation. The SAS code is

```
PROC GENMOD DATA=jones;
  CLASS pid;
  MODEL ecg = time treat /dist = b link = logit;
  REPEATED SUBJECT = pid /TYPE = fixed( 1.00 -0.06
                                        -0.06  1.00 )
                                corrw;
run;
```

The option `TYPE = fixed` calls for a fixed, that is, pre-specified, working correlation matrix. This matrix is a 2 × 2 matrix because there are only two observations per subject in this example. The entries are 1s on the diagonal and -0.06 off the diagonal. The option `corrw` prints the working correlation matrix. The treatment effect is 0.5732 in this case, the 95% CI (0.3628–0.9057), and the p-value equals 0.0171.

Because the working correlation matrix is only of size 2 × 2, GEE working correlation structures yield the same estimated working correlation coefficient, when the correlation coefficient is estimated. Subsequently, the parameter estimates for the treatment effect are identical for these models. This means that the estimates for the exchangeable working correlation, the AR(1) working correlation, and the m-dependent working correlation all are equal. For example, the AR(1) working correlation model is estimated by

```
PROC GENMOD DATA=jones;
  CLASS pid;
  MODEL ecg = time treat /dist = b link = logit;
  REPEATED SUBJECT = pid /TYPE = exch corrw;
run;
```

This model gives an *OR* of 0.5661 with a 95% CI of (0.3588–0.8932). The p-value is 0.0145.

The estimated working correlation coefficient of this model is 0.64, thus extremely high for binary dependent data. The estimate of the working correlation coefficient is substantially different from the crude correlation coefficient of −0.06 which was estimated from the data by ignoring the treatment variable. Nevertheless, the parameter estimates of all models considered in this section were very similar which can be explained by theoretical findings (for references, see, e.g., Ziegler et al. 1998).

Finally, we note that the model considered so far does not include an interaction term between time point and treatment for measuring a carry-over effect. An interaction term can be added by simple using the product sign *.

```
PROC GENMOD DATA=jones;
  CLASS pid;
  MODEL ecg = time treat time * treat
                          /dist = b link = logit;
  REPEATED SUBJECT = pid /TYPE = ind;
run;
```

The p-value for the log *OR* of the carry-over effect appears to be 0.2962 in this example, thus not statistically significant at the 5% test level. Therefore, the carry-over effect is neglected in this model.

35.7 Sample Size Calculations

Sample size and power calculations are especially important for clinical epidemiological studies, where researchers aim at testing a specific pre-specified hypothesis. The simplest approach is to calculate the sample size by ignoring the clustered nature of the data, that is, perform the calculation as if observations were independent. Clustering can then be adjusted for by using the sample size inflation factor (Hsieh et al. 2003). An alternative approach is to use a sample size calculation approach specifically tailored for correlated response data, and a detailed review has been given by Dahmen and Ziegler (2004) (also see Davis 2002, pp. 310; for recent original work, see Ogungbenro et al. 2006; Tu et al. 2004). In a separate paper, we have presented a flexible SAS/IML macro called GEESIZE for sample size calculations (Dahmen et al. 2004) as an extension of the pioneering work by Rochon (1998). The macro is freely available from imbs-luebeck.de. The macro GEESIZE was validated by comparing the computed required sample size with results from closed formulae for two different parallel group study designs. Specifically, binary response variables with unequal cluster sizes were considered for testing a treatment effect, and rates of change in repeated measurements were validated (slope testing) for continuous outcome variables. Results from GEESIZE are also identical to those obtained by closed formulae for crossover designs. With GEESIZE, staggered entry and loss to follow-up can be dealt with, both leading to unequal cluster sizes. The theory underlying GEESIZE and all necessary formulae have been given by Dahmen et al. (2004). Here, we illustrate the use of GEESIZE by recalculating the sample size for the SB-LOT study from Sect. 35.2.3.

Example 35.4. Sample size calculation with GEESIZE
The SB-LOT study described in Sect. 35.2.3 forms the basis of this example. The aim is to demonstrate the edema-protective effect of a vasoactive drug in patients suffering from chronic venous insufficiency after decongestion of the legs. To this end, a double-blind two-arm placebo-controlled 1:1 randomized trial is planned with 5 equidistant follow-ups at weeks 0, 4, 8, 12, and 16 weeks after randomization. Patients either receive the active treatment, that is, medical compression stockings plus SB-LOT, or medical compression stockings plus placebo. The primary endpoint is the continuous variable lower leg volume.

In the original plan of the study, the primary endpoint was the difference of the lower leg volume after the end of the treatment period as compared with baseline. Sample size estimation was based on a standardized difference of 0.4 which was obtained from previous publications. At the one-sided 5% test level α and a power of 0.8 with a dropout rate of 20%, that is, early withdrawals for reasons not related to the efficacy, a sample size of 102 per group was planned (Vanscheidt et al. 2002).

For sample size calculations using GEE, additional assumptions are required. First, the primary hypothesis is now based on slope testing between treatments arms. The link function is the identity link (see _LINK_ in the example code), so that the mean of the dependent variable y_{it} of subject i at time point t given the covariates $\boldsymbol{x}_{it}$ is given by

$$\mathbb{E}(y_{it}|\boldsymbol{x}_{it}) = \mu_{it} = \beta_1 + \beta_2 x_i^{\text{Treat}} + \beta_3 x_{it}^{\text{Time}} + \beta_4 x_i^{\text{Treat}} \cdot x_{it}^{\text{Time}}$$

for the parameters of interest $\beta_1, \ldots, \beta_4$. Because the primary endpoint is quantitative, we use the normal distribution (see _VARI_ in the example code). There are five different time points in total, and the design matrix $\boldsymbol{X}_i$ of a subject i therefore has five rows. Each row consists in four columns: (1) the intercept, (2) the treatment effect is cluster constant and coded as $x_{it} = x_i = 1$ if active treatment, and 0 otherwise, (3) the time point with values from 0 to 4 (see _TIMES_ in the example code), and (4) the treatment by time interaction. The design matrix of a subject in the placebo and the control group are therefore (see _X_ in the example code)

$$\text{Placebo} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 3 & 0 \\ 1 & 0 & 4 & 0 \end{pmatrix}, \qquad \text{Treatment} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 \\ 1 & 1 & 3 & 3 \\ 1 & 1 & 4 & 4 \end{pmatrix}.$$

The hypothesis of interest is

$$H_0 : \beta_4 = 0 \quad \text{versus} \quad H_1 : \beta_4 > 0\,.$$

Because the general hypothesis consists in a single test, the test matrix H has a single row and is equal to $H = (0\;0\;0\;1)$ (see _H_ in the example code).

For the sample size calculation with GEESIZE, we utilize all follow-ups. We assume that no volume difference can be observed in the treatment group over time. However, we assume a linear decrease of the lower leg volume over time. Specifically, the lower leg volume at baseline is assumed to be 2,250 ml in both groups and linearly increases to 2,370 ml at the final follow-up in the placebo group (see _MU_ in the example code). We assume a common standard deviation of 300 ml. These assumptions are consistent with the standardized treatment difference of 0.4 from the published study. Furthermore, we assume a high correlation of $\varrho = 0.9$ (see _RHO_ in the example code) and an AR(1) as true correlation structure (see _PSI_ in the example code). The working correlation is assumed to be the identity matrix; this calculation is automatically carried out. If we assume no dropout, a 5% test level, and a power of 80%, we require 71 patients per group for testing the two-sided hypothesis which is the default in GEESIZE.

If we assume a monotone dropout of 5% for every follow-up, 5.5% of the sample is lost at time point 1, $0.0520 = (1 - 0.055)^2$ is lost at time point 2, $0.0491 = (1 - 0.055)^3$ at time point 3, and 0.0464 at time point 4. Subsequently, 79.75% of the sample is observed at the final follow-up. In the program, these percentages are summarized per group in the matrix _TAU_. Note that these calculations do not include staggered entry which is explained in detail in Dahmen et al. (2004). With these assumptions, we require 97 patients per group. This assumption corresponds to a drop out of approximately 20% at the end of the treatment period. In the original study plan, a one-sided hypothesis was formulated. For a comparison, we also calculate the required sample size using the one-sided hypothesis. If we assume no dropout, we require 56 patients per group, and with a monotone dropout of 5% per visit, we require 76 patients per group. Subsequently, the required sample size would have been lower for the GEE approach.

If the correlation drops, more patients are required. For example, if $\varrho = 0.8$ instead of 0.9, 148 patients would be required per group instead of 97. In contrast, 64 patients per group would be necessary at $\varrho = 0.95$.

```
PROC IML;
%INCLUDE '<PATH>\GeeSize.sas' / NOSOURCE2;
RUN BEGIN;            /* Initialize matrices                 */
_PRINT_ = 1;          /* Detail in print-out                 */
_VARI_ = 'Gaussian';  /* Gaussian variance                   */
_LINK_ = 'Identity';  /* Identity link                       */
_TYPE_I = 0.05;       /* Type I error, two-sided hypothesis  */
_TYPE_II = 0.2;       /* Type II error                       */
_TIMES_ = {0 1 2 3 4};  /* Set of time points                */
_MU_ = { 2250 2280 2310 2340 2370,  /* Placebo group         */
         2250 2250 2250 2250 2250}; /* SB-LOT group          */
_STD_ = 300;          /* Common standard deviation           */
_SNAME_ = { 'Placebo' 'Active treatment' }; /* Labels        */
_X_ = { 1 0 0 0, 1 0 1 0, 1 0 2 0, 1 0 3 0, 1 0 4 0,
                    /* Design matrix in placebo arm          */
        1 1 0 0, 1 1 1 1, 1 1 2 2, 1 1 3 3, 1 1 4 4 };
                    /* Design matrix in active treatment arm */
_H_ = { 0 0 0 1 };  /* Matrix of hypotheses to be tested     */
_PI_ = { 1 1 };     /* Relative group sizes, here: 1:1       */
_TAU_ = { 0.055 0.0520 0.0491 0.0464 0.7975 ,
          0.055 0.0520 0.0491 0.0464 0.7975 }; /* drop out   */
_PSI_ = 1;  /* 0 = CS working correlation, 1 = AR(1)         */
_RHO_= 0.9; /* AR parameter                                  */
RUN GEESIZE;
end ;
end ;
quit;
```

35.8 Illustrative Data Analysis: GEE Analysis for a Parallel Group Design with Repeated Measurements: The SB-LOT Data

In this section, we consider the reanalysis of data from a parallel group randomized controlled trial with repeated measurements. The basic structure and design of this study has been described in Sect. 35.2.3. The primary aim of this study was to demonstrate an edema-protective effect of the vasoactive drug SB-LOT when compared with placebo. The primary analysis was a baseline-adjusted covariance analysis (ANCOVA) for the difference of the lower leg volume at the final visit minus the lower leg volume at baseline for demonstrating a difference of the vasoactive drug when compared with placebo (Vanscheidt et al. 2002). The secondary analysis, which will be presented here, aims at detecting a difference in the slopes by making use of the repeatedness nature of the data. Because the correlation between the measurements at different time points is considered nuisance, this reanalysis represents a typical setting for a GEE analysis.

35.8.1 Descriptive Statistics

In the double-blind placebo-controlled randomized multicenter SB-LOT trial, the primary efficacy endpoint was the lower leg volume measured by water plethysmometry. Patients were randomized to SB-LOT or placebo at baseline. In the first 4 weeks, all patients wore medical compression stockings. At the first follow-up, that is, 4 weeks after randomization, the stockings were discontinued in both treatment groups. Subsequently, the investigators expected an increase in the lower leg volume at subsequent follow-ups in the placebo group, while the lower leg volume was expected to be constant in the active treatment group SB-LOT (see Fig. 35.2). The additional follow-up times were 6, 8, 12, and 16 weeks after randomization. Thus, follow-ups were not equally spaced, and the number of patients is 113 per treatment group. Table 35.4 displays the lower leg volume (ml) in the course of the trial for these patients. The correlation between observations at different time points is high, and the crude correlation estimates are displayed in Table 35.5.

35.8.2 Standard GEE Analysis

For the GEE analysis, we use the lower leg volume at time t as dependent variable y_{it}. Because time points x_{it}^{Time} are not equally spaced in this study, they are

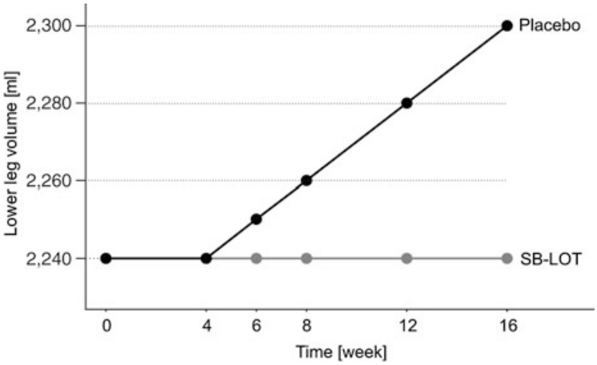


Fig. 35.2 Expected course of the trial. At randomization (week 0), both patients groups are expected to have identical lower leg volumes. Medical compression stockings are discontinued at week 4, and after week 4, the lower leg volume should increase linearly in the placebo group, while it should remain constant in the SB-LOT group

Table 35.4 SB-LOT study: lower leg volume (ml) in the course of the trial

SB-LOT						
Week	0[a]	4	6	8	12	16
n	113	113	113	113	113	113
Mean	2238.3	2235.5	2258.4	2256.7	2261.8	2245.8
Std.Dev.	365.0	371.7	373.4	363.5	369.9	355.4
Min	1409.0	1386.0	1406.0	1404.0	1415.5	1430.5
Max	3443.5	3408.5	3435.0	3408.0	3410.0	3405.0
Placebo						
Week	0[a]	4	6	8	12	16
n	113	113	113	113	113	113
Mean	2260.2	2245.7	2290.1	2294.7	2294.5	2295.9
Std.Dev.	359.0	350.6	350.8	353.4	356.4	354.1
Min	1204.0	1194.5	1234.5	1242.0	1299.5	1246.0
Max	3205.0	3395.5	3304.5	3398.5	3394.0	3268.5

Std.Dev. standard deviation, *Min* minimum, *Max* maximum

[a]Additional use of medical compression stockings

Table 35.5 Estimates of crude correlation coefficients for the SB-LOT data, pooled over both treatment groups

Follow-up	Week 4	Week 6	Week 8	Week 12	Week 16
Week 4	1.0000	0.9594	0.9539	0.9593	0.9403
Week 6	0.9594	1.0000	0.9973	0.9973	0.9971
Week 8	0.9539	0.9973	1.0000	0.9973	0.9973
Week 12	0.9593	0.9973	0.9973	1.0000	0.9973
Week 16	0.9403	0.9971	0.9973	0.9973	1.0000

coded with values according to follow-up weeks starting with $t = 0$ for baseline. Treatment groups have been coded with 1 for placebo and 0 for SB-LOT, that is,

$$x_{it}^{\text{Treat}} = \begin{cases} 1, & \text{if placebo} \\ 0, & \text{if SB-LOT} \end{cases} .$$

Because of the continued use of medical compression stockings, the lower leg volume is expected to be equal after 4 weeks between both treatment arms (see Fig. 35.2). However, the slope should be different in subsequent weeks. Therefore, the model of interest for the mean structure is given by

$$\mathbb{E}(y_{it}) = \beta_1 + \beta_2 x_{it}^{\text{Treat}} + \beta_3 x_{it}^{\text{Time}} + \beta_4 x_{it}^{\text{Treat}} \cdot x_{it}^{\text{Time}}, \quad t = 4, 6, 8, 12, 16 . \quad (35.2)$$

Table 35.6 displays the result from the IEE analysis, and the example code is given below. In the example, `pid` is the person identification date, and the option `type3` requests Type III sum of squares. Because we are interested in slope testing, the parameter of interest Treatment × Time, that is, β_4, is

Table 35.6 Independence estimating equations (IEE) analysis for SB-LOT versus placebo

Parameter	Estimate	Std.Err.	95% CI		z	p
Intercept β_1	2247.21	35.63	2177.37	2317.05	63.06	<0.0001
Treatment β_2	9.48	48.51	−85.60	104.56	0.20	0.8451
Time β_3	0.48	0.93	−1.34	2.31	0.52	0.6045
Treatment × Time β_4	2.51	1.27	0.02	5.00	1.97	0.0484

113 patients per group; *Std.Err.* robust standard error, *CI* confidence interval, z z-score (value of test statistic), p two-sided p-value

displayed in the last row of Table 35.6. Analogously to the ANCOVA model of Vanscheidt et al. (2002), the IEE shows an advantage for SB-LOT over placebo because the slope is significant at the 5% test level ($p = 0.0484$). In greater detail, the difference in the lower leg volume between SB-LOT and placebo increases by 2.51 ml per week, making a difference of 30 ml at the end of the observation period, that is, 12 weeks are discontinuation of medical compression stockings.

```
PROC GENMOD;
  CLASS pid;
  MODEL volume = treat time treat*time
        /dist=normal link=identity type3;
  REPEATED SUBJECT = pid /type= IND;
run;
```

35.8.3 Different Working Correlation Structures

An interesting aspect is whether the model can be substantially improved by using a working correlation matrix which is not the identity matrix. Because measurements are repeated, the natural choice for a working correlation matrix would be an autoregressive working correlation of order 1. However, follow-ups are not equally spaced so that observations at time points 4 and 6 weeks are neighboring as well as observations at time points 8 and 12 weeks. An important assumption in standard GEE packages is that they assume equally spaced follow-ups. Subsequently, choosing AR(1) as working correlation for these data implies that the estimated correlation coefficient can hardly be interpreted. Nevertheless, the estimated working correlation may be used as weight for improving the model fit. Other working correlation matrices may be chosen as well, such as the 1-dependent working correlation or the exchangeable working correlation. The reader should note that even the use of the 1-dependent working correlation does not result in an easily interpretable correlation coefficient because the unequally spaced observations are used for estimating the correlation parameter. Again, the use of a working correlation coefficient might improve the fit, but an interpretation of the estimates is not intended in this example.

Table 35.7 displays the results of several GEE models with different working correlation structures for the slope coefficient β_4, and Table 35.8 gives the estimated

Table 35.7 GEE analysis using various working correlation structures for SB-LOT versus placebo. The parameter estimate (Est.) is the slope coefficient $\hat{\beta}_4$

Correlation	Est.	S.E.	95% CI		z	p	$\hat{\varrho}$
Independence	2.51	1.27	0.02	5.00	1.97	0.0484	–
Autoregressive	2.64	1.21	0.27	5.00	2.19	0.0288	0.9771
Exchangeable	2.51	1.27	0.02	5.00	1.97	0.0484	0.9598
1-dependent	3.06	1.60	−0.07	6.19	1.91	0.0557	0.5773
Unstructured	−4.26	3.72	−11.55	3.03	−1.15	0.2519	Table 35.8

No correlation efficient is estimated in the independence working correlation model
S.E. robust standard error, *CI* confidence interval, z z-score, p two-sided p-value, $\hat{\varrho}$ estimate of the working correlation coefficient

Table 35.8 Estimates of correlation coefficients for the SB-LOT data from the unstructured working correlation model

Follow-up	Week 4	Week 6	Week 8	Week 12	Week 16
Week 4	1.0000	0.9710	0.9482	0.9530	0.9210
Week 6	0.9710	1.0000	0.9850	0.9924	0.9568
Week 8	0.9482	0.9850	1.0000	0.9891	0.9576
Week 12	0.9530	0.9924	0.9891	1.0000	0.9828
Week 16	0.9210	0.9568	0.9576	0.9828	1.0000

correlation coefficients for the unstructured working correlation model from the last row of Table 35.7. Estimates from the IEE and the exchangeable GEE are identical, and estimates from the autoregressive model AR(1) as well as the 1-dependent model are similar in contrast to the estimates from the unstructured working correlation model. The identity of the exchangeable GEE and the IEE can be explained by the theoretical findings of Mancl and Leroux (1996) and Wang and Carey (2003).

The results for the unstructured working correlation model are substantially different. For example, the parameter estimate of the slope coefficient does not agree with the findings of the descriptive statistics (Table 35.4). The results are not reliable, and this can be explained by the instability of the robust covariance matrix. Note that the term robustness refers to the model properties. However, the estimation aspect is a different one. The instability can be investigated by the condition of a covariance matrix which is the ratio of its largest and smallest eigenvalue, and it should not exceed 12.

35.9 Conclusions

In this section, we first discuss the aspect of efficiency of GEE to some extent. Second, we consider several generalizations of the GEE approach. First, the GEE approach does not rely on a correctly specified likelihood function. As there is no free lunch, this robustness property of GEE on the one hand comes at the cost of a loss of efficiency on the other as shown by Chaganty and Joe (2004). The loss in efficiency can be substantial. For example, if the robust variance estimator is used in the standard linear model for a two-arm parallel group study, the GEE variance is twice as large as the standard maximum-likelihood variance estimator; for details, see Ziegler (2011). Because the formulation and presentation of the general results is cumbersome, the reader may refer to Kauermann and Carroll (2001). Therefore, an important question is how the optimal choice of the working correlation matrix can be determined which leads to a minimal loss in efficiency. Furthermore, is it really the efficiency that drives the choice of the working correlation matrix? The simplest choice is the independent working correlation matrix leading to the IEE.

If possible, the investigator should choose a specific working correlation structure for both statistical and biological reasons (Ziegler and Vens 2010). While it is probably intuitive to the reader what is meant by biological plausibility, the statistical reasons for choosing a specific correlation matrix still need to be described. First, using several arguments, the IEE are recommended in clinical epidemiological studies when dropout is not too extreme (Dahmen and Ziegler 2004). Furthermore, as a simple guide, for cluster-type samples, including households or families, the exchangeable working correlation can be used. For longitudinal data, the AR(1) working correlation should be preferred over banded correlation structures (Wang and Carey 2003). Furthermore, m-dependent correlation structures are not biologically plausible. For weakly dependent dichotomous dependent variables, the IEE might be the working correlation structure of choice. An alternative statistical approach is to select the most reasonable working correlation matrix using standard model selection approaches, such as a modification of Akaike's information criterion (AIC), termed quasi-likelihood information criterion (QIC), or bootstrap stability estimates (Cui and Qian 2007; Hin and Wang 2009; Pan 2001a, b; Pan and Connett 2002).

The QIC can also be used for selecting the best mean structure. Alternative approaches for mean structure selection have been considered in several papers. Following the traditional significance testing way, Nuamah et al. (1996) presented an SAS macro for stepwise variable selection. A similar approach is followed by Cantoni (2004). For alternative approaches, the reader may refer to the literature (Pan 2001b; Pan and Connett 2002; Cantoni et al. 2005).

The variance estimator used in GEE is robust against misspecification of the true covariance matrix of the responses. However, it allows valid inferences only if the number of clusters is sufficiently large. In small samples, it has been shown that the null hypothesis is rejected more often than it should; the robust variance estimator is biased and too liberal. This is severe for a very small number of clusters ($n < 20$) but still a problem with as many as 50 clusters. It is worse for unequal cluster sizes than for equal cluster sizes and more extreme for more complex hypotheses. To overcome the liberality and biasedness, a series of modifications has been proposed (for overviews, see, e.g., Dahmen and Ziegler 2004, 2006).

The simple mean is strongly affected by an observation that is far from the rest of the data. In GEE analyses, parameter estimates can also be strongly influenced by outliers which may be clusters, a single cluster or even a single observation. Therefore, the change in parameter estimates is investigated in regression diagnostics if a single observation or a cluster is deleted from the analysis. Corresponding statistics are often called deletion diagnostics. Regression diagnostic techniques that are used in the linear model (Belsley et al. 1980; Cook and Weisberg 1982) or in GLM (Thomas and Cook 1989) can be generalized to the GEE (Vens and Ziegler 2012; Preisser and Perin 2007). Goodness of fit (GOF) statistics for the logistic GEE, including the well-known Tsiatis and Hosmer-Lemeshow type of GOF statistics, have been proposed in a series of different papers, and an excellent summary has been given by Evans and Li (2005).

Because of the clustered nature of the data, two scenarios can be distinguished: the investigation of residuals and deletion diagnostics for one observation and residuals and deletion diagnostics for an entire cluster. Residuals for GEE and the Cook statistic as deletion diagnostic for GEE have been introduced by Ziegler and Arminger (1996). We stress that we have defined the Cook statistic in a way that it does not rely on the correct specification of the covariance matrix. We therefore prefer our regression diagnostics over the definition by Hammill and Preisser (2006) and Venezuela et al. (2007) who use the usual maximum-likelihood-based approach for defining the Cook statistic.

Further regression diagnostics have been proposed, including perturbation methods (Jung 2008). Finally, we stress that the practical computation of the various regression diagnostics is computer time consuming. Therefore, a series of different algorithms has been proposed in recent years (Preisser and Perin 2007; Preisser et al. 2008; Wei and Fung 1999). Standard regression diagnostics are available only in a few standard packages, including a SAS/IML macro (Hammill and Preisser 2006) and `PROC GENMOD` (SAS ver. 9.2).

All analyses discussed so far are applicable to complete data sets or those where data are missing completely at random. This term describes the situation that missing data in the dependent variable y_{it} at time t does not depend on the value of the dependent variable $y_{it'}$ at any other time point t', $t \neq t'$. If, however, systematic differences between complete and incomplete clusters exist and if they are ignored, parameter estimates based on these naïve approaches may be biased and conclusions may be misleading (Rotnitzky and Wypij 1994); for a detailed discussion on missing data, see chapter ▶Missing Data of this handbook.

Although a series of approaches for dealing with missing data in the context of GEE have been proposed, they have rarely been applied in practice because they are hardly available in standard software. Much attention has been given to the situation where some data in the dependent variables from a cluster may be missing, and these are reviewed in Dahmen and Ziegler (2004).

Basically, three different techniques can be distinguished. The first is a two-step method for binary-dependent variables that applies the expectation maximization (EM) algorithm in the first step to obtain unrestricted estimates of multinomial probabilities (Fitzmaurice and Laird 1993). In the second step, the parameter vector of interest is estimated using the estimated response vectors. Two well-known semi-parametric alternatives are imputation methods (Paik 1997) and weighting methods (Robins et al. 1995). The concepts underlying these approaches are intuitive and appealing. Imputation methods replace missing data with estimated values so that a seemingly complete data set emerges. By contrast, weighting methods discard the incomplete data but weigh observations inversely proportional to the probability of observing a response. Imputation and weighting are different, in general, but they yield identical results under specific circumstances (Paik 1997).

Less emphasis has been put on missing covariates. As for missing data in the y, imputation approaches (Xie and Paik 1997a, b) and weighting methods are available (Robins et al. 1994).

In all previous sections, the aim was to consistently estimate the mean structure. The resulting estimating equations are termed GEE in this chapter, and they are usually named GEE1 in the literature. In some applications, the investigator is interested in the estimation of both the mean and the association structure. If both the mean and the association structure are of primary interest, the GEE considered in this chapter can be generalized, and they are usually termed GEE2. For the GEE2 approaches, the reader may refer to the literature (for detailed overviews, see, for example, Liang et al. 1992; Ziegler et al. 1998; Ziegler 2011).

Acknowledgements The author is grateful to Dr. Hans-Heinrich Henneicke-von Zepelin for making the SB-LOT data available and to Dr. Gerlinde Dahmen for analyzing the SB-LOT data.

Appendix 1: Sample Size Inflation Factor

In this appendix, the simple formula of Eq. 35.1 for the sample size inflation factor (SSIF) is derived. For simplicity, we denote $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})'$ for $i = 1, \ldots, n$. The $\boldsymbol{y}_i$ are assumed to be independently identically distributed with element-wise mean $\mathbb{E}(y_{it}) = \mu$, variances given by $\mathbb{V}ar(y_{it}) = \sigma^2$, and covariances given by $\mathbb{C}ov(y_{it}, y_{it'}) = \sigma_{xy}$ for $t = 1, \ldots, T$ and $t \neq t'$. Finally, we let $\mathbb{C}orr(y_{it}, y_{it'}) = \varrho_{xy} = \sigma_{xy} / \sigma^2$.

The simple mean estimator $\bar{y} = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T y_{it}$ is unbiased, and $\mathbb{V}ar(\bar{y})$ reduces to

$$\begin{aligned}
\mathbb{V}ar(\bar{y}) &= \tfrac{1}{n^2T^2} \sum_{i=1}^{n} \mathbb{V}ar\big(\textstyle\sum_{t=1}^T y_{it}\big) \\
&= \tfrac{1}{nT^2}\big(\textstyle\sum_{t=1}^T \mathbb{V}ar(y_{it}) + \tfrac{T(T-1)}{2} 2\mathbb{C}ov(y_{it}, y_{it'})\big) \\
&= \tfrac{1}{nT^2}\big(T\sigma^2 + T(T-1)\sigma_{xy}\big) = \tfrac{1}{nT^2}\big(T\sigma^2 + T(T-1)\sigma^2\varrho_{xy}\big) \\
&= \tfrac{\sigma^2}{nT}\big(1 + (T-1)\varrho_{xy}\big) = \frac{\sigma^2}{nT}\text{SSIF}
\end{aligned}$$

for $t \neq t'$.

Appendix 2: The Roots of the Generalized Estimating Equations (GEE) and Their Statistical Properties

In this section, we consider the roots of the GEE and the statistical properties of the estimator. This section may be ignored by readers who are not interested in technical details. First, we briefly discuss generalized least squares (GLS) and the resulting Aitken estimator, followed by feasible generalized least squares (FGLS). The Aitken estimator based on GLS or FGLS is restricted to continuous dependent variables but allows a general covariance structure. In contrast, the GLM is flexible with respect to the distribution of the responses. These might be continuous, dichotomous, or counts. However, the standard GLM does not allow for correlation between responses. The GEE can be considered a natural combination of both approaches. The GEE are formulated, and their properties are given.

A 2.1 Generalized Least Squares

First, consider the standard linear model for clustered data. Let n be the number of clusters $i = 1, \ldots, n$, and, for simplicity, assume that there are T observations per cluster. The method can easily be generalized to unequal cluster sizes T_i. For each y_{it}, a p dimensional vector of covariates $\boldsymbol{x}_{it}$ is available that possibly contains an intercept. Data are collected in column vectors $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})'$ and the $T \times p$ matrices $\boldsymbol{X}_i = (\boldsymbol{x}'_{i1}, \ldots, \boldsymbol{x}'_{iT})'$. In the context of GEE, the stacked matrix of the $\boldsymbol{X}_i$ is usually termed *X matrix* or *design matrix for the mean structure*, although it need not be a design matrix but may be any matrix of observations.

In this section, we consider modeling of the mean structure of y_{it} given $\boldsymbol{x}_{it}$, and the regression model should allow for correlation within clusters. In the linear model, such an approach is well-known under terms like Aitken estimator or feasible generalized least squares (FGLS). The respective linear model for the mean structure is given by

$$\mu_{it} = \mathbb{E}(y_{it}|\boldsymbol{X}_i) = \mathbb{E}(y_{it}|\boldsymbol{x}_{it}) = \boldsymbol{x}'_{it}\boldsymbol{\beta} = \beta_1 x_{i1t} + \beta_2 x_{i2t} + \ldots + \beta_p x_{ipt}\,, \quad (35.3)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the unknown $p \times 1$ parameter vector of interest. The reader should note that Eq. 35.3 includes the assumption $\mathbb{E}(y_{it}|\boldsymbol{X}_i) = \mathbb{E}(y_{it}|\boldsymbol{x}_{it})$. This means that only the time point specific covariate $\boldsymbol{x}_{it}$ has an effect on the observation y_{it} at time point t. A possible effect of future or past observations is neglected. This assumption has been discussed in detail by Pepe and Anderson (1994) and Pan et al. (2002). Here, we only stress that this assumption is implicitly made in many GEE packages, and it might be violated in family studies. For example, parental smoking might well have an effect on the health status of the entire family, including the offspring.

In the following, we also assume that the covariance matrix of the vector $\boldsymbol{y}_i$ is correctly specified, known, and given by

$$\mathbb{C}ov(\boldsymbol{y}_i|\boldsymbol{X}_i) = \boldsymbol{\Sigma}_i \,. \tag{35.4}$$

Subsequently, the *generalized least squares (GLS) estimator* or *Aitken estimator* $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} = \left(\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\right)^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i \,, \tag{35.5}$$

where $\boldsymbol{X}$ and $\boldsymbol{y}$ are obtained by stacking $\boldsymbol{X}_i$ and $\boldsymbol{y}_i$, and $\boldsymbol{\Sigma}$ is the block diagonal matrix of the matrices $\boldsymbol{\Sigma}_i$. This estimator is based on the score function

$$\boldsymbol{u}(\boldsymbol{\beta}) = \frac{1}{n}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) = \frac{1}{n}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\,,$$

where $\boldsymbol{\mu}_i = (\mu_{i1},\ldots,\mu_{iT})' = \boldsymbol{X}_i\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ is the stacked vector of the $\boldsymbol{\mu}_i$. An estimator $\hat{\boldsymbol{\beta}}$ is obtained by solving $\boldsymbol{u}(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$. The estimator $\hat{\boldsymbol{\beta}}$ is unbiased, and its covariance matrix is given by

$$\mathbb{V}ar\big(\hat{\boldsymbol{\beta}}\big) = \left(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1} = \boldsymbol{A}^{-1}\,. \tag{35.6}$$

$\hat{\boldsymbol{\beta}}$ is the Gauss-Markov estimator, thus it has the smallest variance among all linear unbiased estimators. $\hat{\boldsymbol{\beta}}$ is finite normally distributed if $\boldsymbol{y}_i$ given $\boldsymbol{X}_i$ is normally distributed. If y_i given X_i is not normally distributed, $\hat{\beta}$ is asymptotically normal for $n \to \infty$. The matrix $\boldsymbol{A} = -\mathbb{E}(\partial\boldsymbol{u}(\boldsymbol{\beta})/\partial\boldsymbol{\beta})$ is the *Fisher information matrix.*

Now, we consider the case that estimation is based on the covariance matrix $\boldsymbol{\Sigma}_i$ which has been defined in Eq. 35.4 although

$$\mathbb{V}ar(\boldsymbol{y}_i|\boldsymbol{X}_i) = \boldsymbol{\Omega}_i \neq \boldsymbol{\Sigma}_i \tag{35.7}$$

holds. Equation 35.7 means that $\boldsymbol{\Omega}_i$ is the *true covariance matrix* and that the true covariance matrix is different from the *assumed covariance matrix* $\boldsymbol{\Sigma}_i$ which is used in the linear model. In this case, the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ obtained by Eq. 35.5 remains consistent. The covariance matrix of Eq. 35.6 is, however, no longer appropriate as the assumed covariance matrix $\boldsymbol{\Sigma}_i$, that is, the assumed model, is different from the true one. Subsequently, conclusions-based tests or confidence intervals utilizing the covariance matrix from Eq. 35.6 might be wrong. Instead, one should use the covariance matrix

$$\mathbb{V}ar\big(\hat{\boldsymbol{\beta}}\big) = \left(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1} = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}\,, \tag{35.8}$$

where $\boldsymbol{\Omega}$ is a block diagonal matrix of $\boldsymbol{\Omega}_i$. This estimator is obtained from simple calculations using $\hat{\boldsymbol{\beta}}$ from Eq. 35.5 and using $\mathbb{V}ar(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{\Omega}$. As before, $\hat{\boldsymbol{\beta}}$ is finite normally distributed for normally distributed $\boldsymbol{y}_i$. Otherwise, $\hat{\boldsymbol{\beta}}$ is asymptotically normal for $n \to \infty$. Furthermore, $\boldsymbol{A}^{-1}$ is the inverse Fisher information matrix.

$\boldsymbol{B}$ is termed the *outer product gradient (OPG)* because it can be derived as expected value of the outer product of the gradient $\boldsymbol{u}(\boldsymbol{\beta})$, that is, $\boldsymbol{B}(\boldsymbol{\beta}) = \mathbb{E}\big(\boldsymbol{u}(\boldsymbol{\beta})\boldsymbol{u}(\boldsymbol{\beta})'\big)$.

However, the true covariance matrix of $\hat{\boldsymbol{\beta}}$ is unknown because $\boldsymbol{\Omega}_i$ is unknown. It can, however, be consistently estimated by the so-called "robust estimator of variance" or "sandwich estimator". Here, $\boldsymbol{\Omega}_i$ is replaced by $\hat{\boldsymbol{\Omega}}_i = (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)' = (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})'$. $\hat{\boldsymbol{\Omega}}$ is the diagonal matrix of $\hat{\boldsymbol{\Omega}}_i$. Note that $\hat{\boldsymbol{\Omega}}_i$ is just a replacement for $\boldsymbol{\Omega}_i$, and $\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Omega}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ is a consistent estimator of $\boldsymbol{B}$.

A 2.2 Feasible Generalized Least Squares

Instead of using a pre-specified covariance matrix $\boldsymbol{\Sigma}_i$, one can introduce a model $\boldsymbol{\Sigma}_i(\boldsymbol{\alpha})$ for the covariance matrix which depends on an additional parameter vector $\boldsymbol{\alpha}$. Simple models include an AR(1) structure in case of time-dependent data or an exchangeable structure for families. Given a specific simple structure, the parameter $\boldsymbol{\alpha}$ can be estimated in the first step. In the second step, $\hat{\boldsymbol{\beta}}$ is estimated using $\hat{\boldsymbol{\Sigma}}_i(\hat{\boldsymbol{\alpha}}_i)$. Of course, the structure for the covariance matrix may not involve too many different parameters, and not all covariance structures can be estimated. Only a part is identifiable, thus feasible. The resulting estimator is therefore called *feasible generalized least squares (FGLS) estimator*. A detailed discussion can be found in Greene's excellent textbook (Greene 1993).

A 2.3 Generalized Linear Model

In the linear model, we have been able to allow for within-cluster correlation. However, a linear model often is insufficient for modeling the mean and the variance structure of a problem. The GLM is a well-known alternative, allowing a more flexible modeling of the mean structure than the linear model. The mean structure is given by

$$\mu_{it} = \mathbb{E}(y_{it}|\boldsymbol{x}_{it}) = g(\boldsymbol{x}_{it}'\boldsymbol{\beta}),$$

where g is a non-linear *response function*; g^{-1} is the *link function*. The standard choice in epidemiology for dichotomous responses is the logit link function so that

$$\operatorname{logit}\mathbb{E}(y_{it}|\boldsymbol{x}_{it}) = \operatorname{logit}\mathbb{P}(y_{it}|\boldsymbol{x}_{it}) = \frac{\mathbb{P}(y_{it}|\boldsymbol{x}_{it})}{1 - \mathbb{P}(y_{it}|\boldsymbol{x}_{it})} = \boldsymbol{x}_{it}'\boldsymbol{\beta}\,,$$

which corresponds to the *expit response* function:

$$\mathbb{E}(y_{it}|\boldsymbol{x}_{it}) = \mathbb{P}(y_{it}|\boldsymbol{x}_{it}) = \operatorname{expit}(\boldsymbol{x}_{it}'\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_{it}'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_{it}'\boldsymbol{\beta})}\,.$$

Similarly, the log link is commonly used for count data.

If a distributional assumption for the responses is made, the variance function is automatically determined. In general, the variance $v_{it} = \mathbb{V}ar(y_{it}|\boldsymbol{x}_{it})$ is given by $v_{it} = h(\mu_{it}) \cdot \phi$. A common assumption for binary responses is the binomial distribution, yielding $v_{it} = \mu_{it}(1 - \mu_{it})$, thus $\phi = 1$. For count data, the standard choice is a Poisson distribution, where the mean and the variance are equal. Subsequently, $v_{it} = \mu_{it}$, and the parameter $\phi = 1$. In contrast, for the normal distribution, the variance is given by σ^2, thus independent of the mean. In this case, $v_{it} = 1 \cdot \phi = \sigma^2$ for $\phi = \sigma^2$.

If all observations y_{it} are independent, the parameter vector $\boldsymbol{\beta}$ can be estimated using standard maximum likelihood (ML). Specifically, the distributional assumption determines the score equations which are the first derivative of the log-likelihood function with respect to $\boldsymbol{\beta}$. For n-independent observations y_i, $i = 1, \ldots, n$, the score function has the form

$$\boldsymbol{u}(\boldsymbol{\beta}) = \frac{1}{n}\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{D}_i'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)\,. \tag{35.9}$$

Here, $\boldsymbol{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}'$ is the matrix of first derivatives and $\boldsymbol{\Sigma}_i$ is the diagonal matrix of variances $\boldsymbol{\Sigma}_i = \text{diag}(v_{it})$. Subsequently, $\boldsymbol{D}$ is the stacked matrix of the $\boldsymbol{D}_i$ and $\hat{\cdot}$ indicates the estimates. In Eq. 35.9, the observations y_{it} are explicitly assumed to be independent because the variance matrix $\boldsymbol{\Sigma}_i$ is diagonal. The estimating equations $\boldsymbol{u}(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$ are therefore termed independence estimating equations (IEE). Except for the linear model, Eq. 35.9 generally has no closed formula solution, and the estimator is determined via iterative approaches like Fisher scoring or iterative re-weighted least squares (Dennis and Schnabel 1983). The estimator $\hat{\boldsymbol{\beta}}$ is asymptotically normal with mean $\boldsymbol{\beta}$ and covariance matrix $\mathbb{V}ar(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\right)^{-1} = \boldsymbol{A}^{-1}$ which can be estimated consistently by

$$\widehat{\mathbb{V}ar}(\hat{\boldsymbol{\beta}}) = \left(\hat{\boldsymbol{D}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{D}}\right)^{-1} = \left(\sum_{i=1}^{n}\hat{\boldsymbol{D}}_i\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{D}}_i\right)^{-1}\,. \tag{35.10}$$

A 2.4 Generalized Estimating Equations: Formulation and Properties

In the last section, we assumed independence of observations y_{it} within cluster i. This assumption is most likely inadequate. The dependence was, however, only considered in the GLS and the FGLS models. We can now combine the Aitken estimator approach with the GLM method. If observations y_{it} within cluster i are correlated, then the true covariance matrix $\boldsymbol{\Omega}_i$ is not diagonal. Zeger et al. (1985) therefore proposed to use the robust estimator of variance $\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}$ instead of the inverse Fisher information matrix $\boldsymbol{A}$:

$$\widehat{\mathbb{V}ar}(\hat{\boldsymbol{\beta}}) = \left(\hat{\boldsymbol{D}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{D}}\right)^{-1}\left(\hat{\boldsymbol{D}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{D}}\right)\left(\hat{\boldsymbol{D}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{D}}\right)^{-1} \tag{35.11}$$
$$= \left(\sum_{i=1}^{n}\hat{\boldsymbol{D}}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{D}}_i\right)^{-1}\left(\sum_{i=1}^{n}\hat{\boldsymbol{D}}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\Omega}}_i\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{D}}_i\right)\left(\sum_{i=1}^{n}\hat{\boldsymbol{D}}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{D}}_i\right)^{-1}.$$

In contrast to Eq. 35.10, Eq. 35.11 is a consistent estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$ even if the correlations are dependent within a cluster. However, the estimator will not be very efficient in many cases because estimates are based on a model which assumes independence of observations. Liang and Zeger (1986) and Zeger and Liang (1986) proposed to increase the efficiency by a synthesis of FGLS and GLM. First, the mean structure μ_{it} and the variance function v_{it} are specified as in standard GLM. Second, a covariance matrix $\boldsymbol{\Sigma}_i$ that need not be a diagonal matrix is assumed as in FGLS. In practice, the covariance matrix is estimated in a first step, yielding $\hat{\boldsymbol{\Sigma}}_i$. This covariance matrix need not be correctly specified and is therefore termed "working covariance matrix." The resulting score function is given by

$$\boldsymbol{u}(\boldsymbol{\beta}) = \frac{1}{n}\boldsymbol{D}'\hat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}\right) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{D}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\left(\boldsymbol{y}_i-\boldsymbol{\mu}_i\right), \tag{35.12}$$

which needs to be set to $\mathbf{0}$ and solved for $\boldsymbol{\beta}$ to obtain the GEE estimator $\hat{\boldsymbol{\beta}}$. The working covariance matrix should be as close to the probably unknown true covariance matrix $\boldsymbol{\Omega}_i$. However, it is not the aim to have an interpretable covariance matrix. It is only used as weight for increasing the efficiency of the estimator.

The properties summarized in Box 35.2 hold for an estimator $\hat{\boldsymbol{\beta}}$ solving the GEE which are given by $\boldsymbol{u}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$.

Box 35.2. Properties of the GEE estimator $\hat{\beta}$

1. *Existence.* There exists asymptotically a GEE estimator $\hat{\boldsymbol{\beta}}$ for the true parameter vector.
2. *Strong consistency.* The GEE estimator converges almost surely to the true parameter vector.
3. *Asymptotic unbiasedness.* The GEE estimator is asymptotically unbiased, that is, $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
4. *Asymptotic normality.* The GEE estimator $\hat{\boldsymbol{\beta}}$ is asymptotically normal. More specifically, $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed with mean $\boldsymbol{\beta}$ and robust covariance matrix $\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1} = \boldsymbol{C}^{-1}$.
5. *Estimation of covariance matrix.* Strongly consistent estimators of the Fisher information matrix $\boldsymbol{A}$ and the OPG $\boldsymbol{B}$ are, for example, given by

$$\hat{\boldsymbol{A}}(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{A}}_i = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{D}}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{D}}_i\right) \quad \text{and}$$

$$\hat{\boldsymbol{B}}(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{B}}_i = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{D}}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\Omega}}_i\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{D}}_i\right),$$

where $\hat{\boldsymbol{D}}_i$, $\hat{\boldsymbol{\mu}}_i$, and $\hat{\boldsymbol{\Sigma}}_i$ are the estimators of $\boldsymbol{D}_i$, $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$, respectively, and $\hat{\boldsymbol{\Omega}}_i = (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)'$.

6. *Efficiency.* If the working covariance matrix $\boldsymbol{\Sigma}_i$ is correctly specified, that is, if $\boldsymbol{\Sigma}_i = \boldsymbol{\Omega}_i$, then the GEE estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is asymptotically efficient.

A few remarks are required after having formulated these strong statements. First, they are only valid under suitable regularity conditions. For example, necessary for the strong consistency is that the score function (35.12) is derived from a distribution belonging to the linear exponential family (Gourieroux et al. 1984). An important additional assumption is that the mean structure $\boldsymbol{\mu}_i = g(\boldsymbol{x}_i'\boldsymbol{\beta})$ is correctly specified. If it is misspecified, $\boldsymbol{\beta}$ cannot be consistently estimated, and $\hat{\boldsymbol{\beta}}$ does not converge to the true parameter vector $\boldsymbol{\beta}$. Furthermore, a covariance matrix $\boldsymbol{\Omega}_i$ needs to exist. Otherwise, estimation of the covariance would not be possible.

An interesting aspect is that the estimator is consistent, although a two-step approach is taken for estimation. To repeat, the working covariance matrix is estimated first. In the second step, the estimating equations for the mean structure are solved. In fact, it can be shown using properties 8.5 and 8.19 of Gourieroux and Monfort (1995) that the GEE estimator $\hat{\boldsymbol{\beta}}$ obtained in the two-step approach fulfills these assumptions. Alternatively, the proof can be obtained using the generalized method of moments (GMM) (Ziegler 1995).

Above, we have considered specifying a working covariance matrix. However, given the variance function, specification of a working correlation matrix is sufficient. In detail, both the mean μ_{it} and the variance v_{it} of y_{it} given $\boldsymbol{x}_{it}$ have been specified in the GLM. For GEE, the correct specification of the variance v_{it} is not required for consistent estimation of $\boldsymbol{\beta}$, and this is in contrast to GLM. However, with the specific choice of the variance according to a GLM, it is not necessary to specify the entire working covariance matrix. Instead, the standard decomposition

$$\varrho_{itt'} = \mathbb{C}orr(y_{it}, y_{it'}) = \frac{\mathbb{C}ov(y_{it}, y_{it'})}{\sqrt{\mathbb{V}ar(y_{it})\,\mathbb{V}ar(y_{it'})}} = \frac{\sigma_{itt'}}{\sqrt{v_{it}\,v_{it'}}}$$

can be used for simplification. Specifically, the working covariance matrix $\boldsymbol{\Sigma}_i$ can be obtained as

$$\boldsymbol{\Sigma}_i = \boldsymbol{V}_i^{1/2}\boldsymbol{R}_i\boldsymbol{V}_i^{1/2},$$

where $\boldsymbol{V}_i = \text{diag}(v_{it})$ is the diagonal matrix of variances v_{it} in cluster i, and $\boldsymbol{R}_i = \mathbb{C}orr(\boldsymbol{y}_i|\boldsymbol{X}_i)$ is the working correlation matrix of $\boldsymbol{y}_i$ given $\boldsymbol{X}_i$. $\boldsymbol{R}_i$ needs to be a positive definite correlation matrix that – in principle – should be as close to the true correlation matrix as possible.

References

Ballinger GA (2004) Using generalized estimating equations for longitudinal data analysis. Organ Res Method 7:127–150

Baradat P, Maillart M, Marpeau A, Slak MF, Yani A, Pastiszka P (1996) Utility of terpenes to assess population structure and mating patterns in conifers. In: Philippe B, Thomas A, Müller-Starck G (eds) Population genetics and genetic conservation of forest trees. Academic Publishing, Amsterdam, pp 5–27

Belsley DA, Kuh E, Welsch RE (1980) Regression diagnostics: identifying influential data and sources of collinearity. Wiley, New Nork

Cantoni E (2004) A robust approach to longitudinal data analysis. Can J Statist 32:169–180

Cantoni E, Flemming JM, Ronchetti E (2005) Variable selection for marginal longitudinal generalized linear models. Biometrics 61:507–514

Chaganty N, Joe H (2004) Efficiency of generalized estimating equations for binary responses. J R Stat Soc B 66:851–860

Cochran WG (1963) Sampling techniques, 2nd edn. Wiley, New York

Cook RD, Weisberg S (1982) Residuals and influence in regression. Chapman and Hall, New York

Cui J, Qian G (2007) Selection of working correlation structure and best model in GEE analyses of longitudinal data. Commun Stat Simul Comput 36:987–996

Dahmen G, Ziegler A (2004) Generalized estimating equations in controlled clinical trials: hypotheses testing. Biom J 46:214–232

Dahmen G, Ziegler A (2006) Independence estimating equations for controlled clinical trials with small sample size: interval estimation. Methods Inf Med 45:430–434

Dahmen G, Rochon J, König IR, Ziegler A (2004) Sample size calculations for controlled clinical trials using generalized estimating equations (GEE). Methods Inf Med 43:451–456

Davis CS (2002) Statistical methods for the analysis of repeated measurements. Springer, New York

Dennis J, Schnabel R (1983) Numerical methods for unconstrained optimization and nonlinear equations. Prentice-Hall, Englewood Cliffs

Diggle PJ, Liang KY, Zeger SL (1994) Analysis of longitudinal data. Clarendon Press, Oxford

Dobson AJ (2001) Introduction to generalized linear models, 2nd edn. Chapman and Hall, London

Evans S, Li L (2005) A comparison of goodness of fit tests for the logistic GEE model. Stat Med 24:1245–1261

Fahrmeir L, Pritscher L (1996) Regression analysis of forest damage by marginal models for correlated ordinal responses. Environ Ecol Stat 3:257–268

Fahrmeir L, Tutz G (1994) Multivariate statistical modelling based on generalized linear models. Springer, New York

Fitzmaurice GM, Laird NM (1993) A likelihood-based method for analysing longitudinal binary responses. Biometrika 80:141–151

Gail MH, Wieand S, Piantadosi S (1984) Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika 71:431–444

Gourieroux C, Monfort A (1995) Statistics and econometric models, vol 1. Cambridge University Press, Cambridge

Gourieroux C, Monfort A, Trognon A (1984) Pseudo maximum likelihood methods: theory. Econometrics 52:681–700

Greene W (1993) Econometric analysis, 2nd edn. Macmillan, New York

Hammill BG, Preisser JS (2006) A SAS/IML software program for GEE and regression diagnostics. Comput Stat Data Anal 51:1197–1212

Hanley JA, Negassa A, Edwardes MD (2000) GEE analysis of negatively correlated binary responses: a caution. Stat Med 19:715–722

Hanley JA, Negassa A, Edwardes MD, Forrester JE (2003) Statistical analysis of correlated data using generalized estimating equations: an orientation. Am J Epidemiol 157:364–375

Hasturk H, Nunn M, Warbington M, Van Dyke TE (2004) Efficacy of a fluoridated hydrogen peroxide-based mouthrinse for the treatment of gingivitis: a randomized clinical trial. J Periodontol 75:57–65

Hin LY, Wang YG (2009) Working-correlation-structure identification in generalized estimating equations. Stat Med 28:642–658

Hsieh FY, Lavori PW, Cohen HJ, Feussner JR (2003) An overview of variance inflation factors for sample-size calculation. Eval Health Prof 26:239–257

Jones B, Kenward MG (1989) Design and analysis of cross-over trials. Chapman & Hall, London

Jones B, Kenward MG (2003) Design and analysis of cross-over trials, 2nd edn. Chapman & Hall, London

Jung KM (2008) Local influence in generalized estimating equations. Scand J Stat 35:286–294

Kauermann G, Carroll RJ (2001) A note on the efficiency of sandwich covariance matrix estimation. J Am Stat Assoc 96:1387–1396

Lechner M, Lollivier S, Magnac T (2008) Parametric binary choice models. In: Mátyás L, Sevestre P (eds) The econometrics of panel data, 3rd edn. Springer, Heidelberg, pp 215–245

Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22

Liang K-Y, Zeger SL, Qaqish B (1992) Multivariate regression analysis for categorical data. J R Stat Soc B 54:3–40

Mancl LA, Leroux BG (1996) Efficiency of regression estimates for clustered data. Biometrics 52:500–511

Martus P, Stroux A, Jünemann AM, Korth M, Jonas JB, Horn FK, Ziegler A (2004) GEE approaches to marginal regression models for medical diagnostic tests. Stat Med 23: 1377–1398

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall, London

Nelder JA, Wedderburn RW (1972) Generalized linear models. J R Stat Soc A 135: 370–384

Nuamah IF, Qu Y, Amini SB (1996) A SAS macro for stepwise correlated binary regression. Comput Method Program Biomed 49:199–210

Ogungbenro K, Aarons L, Graham G (2006) Sample size calculations based on generalized estimating equations for population pharmacokinetic experiments. J Biopharm Stat 16: 135–150

Paik MC (1997) The generalized estimating equation approach when data are not missing completely at random. J Am Stat Assoc 92:1320–1329

Pan W (2001a) Akaike's information criterion in generalized estimating equations. Biometrics 57:120–125

Pan W (2001b) Model selection in estimating equations. Biometrics 57:529–534

Pan W, Connett JE (2002) Selecting the working correlation structure in generalized estimating equations with application to the lung health study. Stat Sin 12:475–490

Pan W, Louis TA, Connett JE (2002) A note on marginal linear regression with correlated response data. Am Stat 54:191–195

Pepe MS, Anderson GL (1994) A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. Commun Stat Simul Comput 23:939–951

Preisser JS, Perin J (2007) Deletion diagnostics for marginal mean and correlation model parameters in estimating equations. Stat Comput 17(4):381–393. doi:10.1007/s11222-007-9031-1

Preisser JS, Qaqish BF, Perin J (2008) A note on deletion diagnostics for estimating equations. Biometrika 95:509–513

Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 89:846–866

Robins JM, Rotnitzky A, Zhao LP (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc 90:106–120

Rochon J (1998) Application of GEE procedures for sample size calculations in repeated measures experiments. Stat Med 17:1643–1658

Rotnitzky A, Wypij D (1994) A note on the bias of estimators with missing data. Biometrics 50:1163–1170

Ryan L (1992) The use of generalized estimating equations for risk assessment in developmental toxicity. Risk Anal 12:439–447

Stokes ME (1999) Recent advances in categorical data analysis. Paper presented at the 24th annual meeting of the SAS users group international conference, Miami Beach. http://support.sas.com/rnd/app/papers/abstracts/categorical.html

Tan AG, Mitchell P, Burlutsky G, Rochtchina E, Kanthan G, Islam FM, Wang JJ (2008) Retinal vessel caliber and the long-term incidence of age-related cataract: the Blue Mountains Eye Study. Ophthalmology 115:1693–1698

Thomas W, Cook RD (1989) Assessing influence on regression coefficients in generalized linear models. Biometrika 76:741–749

Tu XM, Kowalski J, Zhang J, Lynch KG, Crits-Christoph P (2004) Power analyses for longitudinal trials and other clustered designs. Stat Med 23:2799–2815

Vanscheidt W, Rabe E, Naser-Hijazi B, Ramelet AA, Partsch H, Diehm C, Schultz-Ehrenburg U, Spengel F, Wirsching M, Götz V, Schnitker J, Henneicke-von Zepelin HH (2002) The efficacy and safety of a coumarin-/troxerutin-combination (SB-LOT) in patients with chronic venous insufficiency: a double blind placebo-controlled randomised study. VASA 31: 185–190

Venezuela MK, Botter DA, Sandoval MC (2007) Diagnostic techniques in generalized estimating equations. J Stat Comput Simul 77:879–888

Vens M, Ziegler A (2012) Generalized estimating equations and regression diagnostics for longitudinal controlled clinical trials: A case study. Comput Stat Data Anal 56(5):1232–1242. doi:10.1016/j.csda.2011.04.010

Wang Y-G, Carey V (2003) Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance. Biometrika 90:1–24

Wei WH, Fung WK (1999) The mean-shift outlier model in general weighted regression and its applications. Comput Stat Data Anal 30:429–441

Xie F, Paik MC (1997a) Generalized estimating equation model for binary outcomes with missing covariates. Biometrics 53:1458–1466

Xie F, Paik MC (1997b) Multiple imputation methods for the missing covariates in generalized estimating equation. Biometrics 53:1538–1546

Yang J, Peek-Asa C, Jones MP, Nordstrom DL, Taylor C, Young TL, Zwerling C (2008) Smoke alarms by type and battery life in rural households. Am J Prev Med 35:20–24

Zeger SL, Liang KY (1986) Longitudinal data analysis for discrete and continuous outcomes. Biometrics 42:121–130

Zeger S, Liang K, Self S (1985) The analysis of binary longitudinal data with time-independent covariates. Biometrika 72:31–38

Ziegler A (1995) The different parameterizations of the GEE1 and the GEE2. In: Seeber GUH, Francis BJ, Hatzinger R, Steckel-Berger G (eds) Statistical modelling proceedings of the 10th international workshop on statistical modelling. Lecture Notes in statistics, vol 104. Springer, Heidelberg, pp 315–324

Ziegler A, Arminger G (1996) Parameter estimation and regression diagnostics using generalized estimating equations. In: Faulbaum F, Bandilla W (eds) SoftStat '95. Advances in statistical software 5. Lucius & Lucius, Heidelberg, pp 229–237

Ziegler A, Kastner C, Blettner M (1998) The generalised estimating equations: an annotated bibliography. Biom J 40:115–139

Ziegler A, Kastner C, Brunner D, Blettner M (2000) Familial associations of lipid profiles: a generalized estimating equations approach. Stat Med 19:3345–3357

Ziegler A, Kastner C, Chang-Claude J (2003) Analysis of pregnancy and other factors on detection of human papilloma virus (HPV) infection using weighted estimating equations for follow-up data. Stat Med 22:2217–2233

Ziegler A, Vens M (2010) Generalized estimating equations: Notes on the choice of the working correlation matrix. Methods Inf Med 49(5):421–425. doi:10.3414/ME10-01-0026

Ziegler A (2011) Generalized estimating equations: Theory. Springer, New York.

Meta-Analysis in Epidemiology 36

Maria Blettner, Ulrike Krahn, and Peter Schlattmann

Contents

M. Blettner (✉) • U. Krahn
Institute for Medical Biostatistics, Epidemiology and Informatics, Johannes Gutenberg University, Mainz, Germany

P. Schlattmann
Institute of Medical Statistics, Computer Sciences and Documentation, Jena University Hospital, Jena, Germany

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_21,

36.1 Introduction

The use of meta-analyses in order to synthesize the evidence from epidemiological studies has become more and more popular recently. It has been estimated by Egger et al. (1998) that from articles retrieved by MEDLINE with the medical subject heading (MeSH) term "meta-analysis" some 33% reported results of a meta-analysis from randomized clinical trials and nearly the same proportion (27%) were from observational studies, including 12% papers in which the etiology of a disease was investigated. The remaining papers include methodological publications or review articles. Reasons for the popularity of meta-analyses are the growing information in the scientific literature and the need of timely decisions for risk assessment or in public health. Methods for meta-analyses in order to summarize or synthesize evidence from randomized controlled clinical trials have been continuously developed during the last years. In 1993, the Cochrane Collaboration was established as an international organization, which provides systematic reviews to evaluate healthcare interventions. They have published a handbook (Higgins and Green 2009) with detailed information on how to conduct systematic reviews of randomized clinical trials. While methods for meta-analyses of randomized clinical trials are now also summarized in several text books, for example, Sutton et al. (2000) and Whitehead (2002), and in a handbook by Egger et al. (2001a) and Dickersin (2002) argued that statistical methods for meta-analyses of epidemiological studies are still behind in comparison to the progress that has been made for randomized clinical trials. The use of meta-analyses for epidemiological research caused many controversial discussions; see, for example, Blettner et al. (1999), Berlin (1995), Greenland (1994), Feinstein (1995), Olkin (1994), Shapiro (1994a, b), or Weed (1997) for a detailed overview of the arguments. The most prominent arguments against meta-analyses are the fundamental issues of confounding, selection bias, as well as the large variety and heterogeneity of study designs and data collection procedures in epidemiological research. Despite these controversies, results from meta-analyses are often cited and used for decisions. They are often seen as the fundamentals for risk assessment. They are also performed to summarize the current state of knowledge often prior to designing new studies.

This chapter will first describe reasons for meta-analyses in epidemiological research and then illustrate how to perform a meta-analysis with the focus on meta-analysis of published data. Furthermore, network meta-analyses are introduced, which are a new methodological tool.

36.2 Reasons for Meta-Analysis in Epidemiology

One major issue in assessing causality in epidemiology is "consistency" as pointed out by Hill in 1965. The extent to which an observed association is similar in different studies, with different designs, using different methods of data collection and exposure assessment, by different investigators, and in different regions or countries is an essential criterion for causality. If different studies with inconsistent results are known, there is a need for understanding the differences. Reasons may be small sample sizes of individual studies (chance), different methods of exposure assessment (measurement errors), different statistical analyses (e.g., adjustment for confounding), or the use of different study populations (selection bias). Also, Thompson et al. (1997) showed that different baseline risks may cause heterogeneity. The goal of a meta-analysis is then to investigate whether the available evidence is consistent and/or to which degree inconsistent results can be explained by random variation or by systematic differences between design, setting, or analysis of the study as has been pointed out by Weed (2000).

Meta-analyses are often performed to obtain a combined estimator of the quantitative effect of the risk factor such as the relative risk (*RR*) or the odds ratio (OR). As single studies are often far too small to obtain reliable risk estimates, the combination of data of several studies may lead to more precise effect estimates and increased statistical power. This is mainly true if the exposure leads only to a small increase (or decrease) in risk or if the disease or the exposure of interest is rare. One example is the risk of developing lung cancer after the exposure to passive smoking where relative risk estimates in the order of 1.2 have been observed; see Boffetta (2002) for a summary of the epidemiological evidence. Another typical example is the association between childhood leukemia and exposure to electromagnetic fields. Meinert and Michaelis (1996) have performed a meta-analysis of the available case-control studies as the results of the investigations were inconsistent. Although many huge case-control studies have been performed in the last decade, in each single study, only a few children were categorized as "highly exposed." In most publications, a small but non-significant increase in risk was found, but no single study had enough power to exclude that there is no association between EMF exposure and childhood leukemia.

Sometimes, meta-analyses are also used to investigate more complex dose-response functions. For example, Tweedie and Mengersen (1995) investigated the dose-response relationship of exposure to passive smoking and lung cancer. A meta-analysis was also undertaken by Longnecker et al. (1988) to study the dose-response of alcohol consumption and breast cancer risk. However, results were limited as not enough data were present in several of the included publications. Interestingly, a large group of investigators led by Hamajima et al. (2002) has recently used individual patient data from 53 studies including nearly 60,000 cases for a reanalysis. It has been shown by Sauerbrei et al. (2001) in a critique that meta-analysis from aggregated data may be too limited to perform a dose-response analysis. A major limitation is that different categories are used in different publications. Thus, dose-response analyses are restricted to published values. Meta-analyses of published data have their main merits for exploring heterogeneity between studies and to provide crude quantitative estimates but probably less for investigating complex dose-response relationships.

36.3 Different Types of Overviews

Approaches for summarizing evidence include five different types of overviews: first, traditional narrative reviews that provide a qualitative but not a quantitative assessment of published results. Methods and guidelines for reviews have been recently published by Weed (1997).

Second, meta-analyses from literature (MAL) which are generally performed from freely available publications without the need of cooperation and without agreement of the authors from the original studies. They are comparable to a narrative review in many respects but include quantitative estimate(s) of the effect of interest. One recent example is a meta-analysis by Zeeger et al. (2003) of studies investigating some familial clustering of prostate cancer. Another meta-analysis has been recently published by Allam et al. (2003) on the association between Parkinson disease, smoking, and family history.

Third, meta-analyses with individual patient data (MAP) in which individual data from published and sometimes also unpublished studies are reanalyzed. Often, there is a close cooperation between the researcher performing the meta-analysis and the investigators of the individual studies. The new analysis may include specific inclusion criteria for patients and controls, new definition of the exposure and confounder variables, and new statistical modeling. This reanalysis may overcome some but not all of the problems of meta-analyses of published data (Blettner et al. 1999). They have been performed in epidemiological research for many years. One of the largest investigations of this form was a recent investigation on breast cancer and oral contraceptive use, where data from 54 case-control studies were pooled and reanalyzed (CGHFBC 1996). A further international collaboration led by Lubin and colleagues were set up to reanalyze data from 11 large cohort studies on lung cancer and radon among uranium miners. The reanalysis allowed a refined dose-response analysis and provided data for radiation protection issues. Pooled reanalyses are mostly performed by combining data from studies of the same type only. For example, Hung et al. (2003) reanalyzed data from all case-control studies in which the role of genetic polymorphisms for lung cancer in non-smokers was investigated. The role of diet for lung cancer was recently reviewed by Smith-Warner et al. (2002) in a qualitative and quantitative way by combining cohort studies. An overview of methodological aspects for a pooled analysis of data from cohort studies was published by Bennett (2003). There are methods under development which allow to combine individual and aggregate data. See, for example, the work by Riley et al. (2007) or Sutton et al. (2008).

Fourth, prospectively planned pooled meta-analyses of several studies in which pooling is already a part of the protocol. Data collection procedures, definitions of variables are as far as possible standardized for the individual studies. The statistical analysis has many similarities with the meta-analysis based on individual data. A major difference, however, is that joint planning of the data collection and analysis increase the homogeneity of the included data sets. However, in contrast to multicenter randomized clinical trials, important heterogeneity between the study centers still may exist. This heterogeneity may arise from differences in populations, in the relevant confounding variables (e.g., race may only be a confounder in some centers) and potentially differences in ascertainment of controls. For example, complete listings of population controls are available in some but not all countries. In the latter situation sometimes neighborhood controls are used. Mainly in occupational epidemiology, those studies are rather common, many of them were initiated by international bodies such as the International Agency for Research on Cancer (IARC) as the international pooled analysis by Boffetta et al. (1997) of cancer mortality among persons exposed to man-made mineral fiber. Another example for a prospectively planned pooled meta-analysis is given by a large brain tumor study initiated by the IARC including data from eight different countries (see Schlehofer et al. 1999).

Fifth, network meta-analyses which is a method that is increasingly used to estimate comparative effectiveness of treatments not compared directly in randomized controlled trials. This technique is introduced in more detail in Sect. 36.7.

Steinberg et al. (1997) compared the effort required and the results obtained of MAL and MAP with an application to ovarian cancer. Certainly, MAL are easier to perform, cheaper, and faster than MAP. Their credibility may be more questionable as discussed by many authors; see, for example, Blettner et al. (1999) or Egger et al. (1998). Statistical issues of pooling data from case-control studies have been investigated by Stukel et al. (2001) recently. The authors proposed a two-step approach and showed conditions under which the two-step approach gives similar results in comparison to the pooled analysis including all data. Here, the two-step approach implies to estimate first the odds ratio for each study in the usual way. Then in the second step, a combined estimator using either a fixed or random effects model is calculated.

36.4 Steps in Performing a Meta-Analysis

Each type of overview needs a clear study protocol that describes the research question and the design, including how studies are identified and selected, the statistical methods to use, and how the results will be reported. This protocol should also include the exact definition of the disease of interest, the risk factors, and the potential confounding variables that have to be considered. In accordance with Friedenreich (1993) and Jones (1992), the following steps are needed for a meta-analysis/pooled analysis.

Step 1. Define a clear and focused topic for the review: As for any other investigation, a clear protocol in which the research hypothesis, that is, the objectives of the meta-analysis are described, is mandatory. This protocol should include the exact definition of the disease of interest, the risk factors, and the potential confounding variables that have to be considered. The protocol should also include details on the steps that are described below, including specification of techniques for location of the studies, the statistical analysis, and the proposed publications.

Step 2. Establish inclusion and exclusion: It is important to define in advance which studies should be included into the meta-analysis. These criteria may include restrictions on the publication year as older studies may not be comparable to newer ones, on the design of the investigation, for example, to exclude ecological studies. Friedenreich (1994) has also proposed quality criteria to evaluate each study. Whether these criteria, however, should be used as inclusion criteria is discussed controversially. Another decision is whether studies that are only published as abstracts or internal communications should be included (Cook et al. 1993). A rule for the inclusion or exclusion of papers with repeated publication of the data is required. For example, for cohort studies, often several publications with different follow-up periods can be found. As one out of many examples, a German study among rubber workers by Straif et al. (1999, 2000) can be mentioned. In one paper, 11,633 workers were included, while the second paper is based on a subcohort of only 8,933 persons. Which results are more appropriate for the meta-analysis?

Step 3. Locate all studies (published and unpublished) that are relevant to the topic: Since the existence of electronic databases, retrieval of published studies has become much easier. Mainly systems like MEDLINE or CANCERLIT from the National Library of Medicine are valuable sources to locate publications. However, as Dickersin et al. (1994) showed for some examples as little as 50% of the publications were found by electronic searches. Therefore, there is a need to extend the search by manual checks of the reference lists of retrieved papers, monographs, books, and if possible by personal communications with researchers in the field. A clear goal of the search has to be to identify all relevant studies on the topic that meet the inclusion criteria. Egger et al. (2003) have pointed out that the completeness of the literature search is an important feature of the meta-analysis to avoid publication bias or selection bias. Of course, the publication should include the search strategies as well as the keywords and the databases used for electronic searches.

Step 4. Abstract information from the publications: The data collection step in a meta-analysis needs as much care as in other studies. In the meta-analysis, the unit of observation is the publication, and defined variables have to be abstracted from the publication (Stock 1995). In epidemiological studies, the key parameter is often the relative risk or odds ratio. Additionally, standard error, sample size, treatment of confounders, and other characteristics of the study design and data collection procedure need to be abstracted to assess the quality of the study. This is also important for subgroup analyses or for a sensitivity analysis. An abstract form has to be created before abstracting data. This form should be tested like other instruments in a pilot phase. Unfortunately, it may not always be possible to abstract the required estimates directly, for example, standard errors are not presented and have to be calculated based on confidence intervals (Greenland 1987). It may be necessary to contact the investigators to obtain further information if results are not published in sufficient detail. Abstracting and classification of study characteristics is the most time-consuming part of the meta-analysis. It has been recommended to blind the data abstractors, although some authors argue that blinding may not have a major influence on the results; for further discussion, see Berlin et al. (1997). Additionally, the rater may be acquainted with some of the studies and blinding cannot be performed. Another requirement is that two persons should perform the abstraction in parallel. When a meta-analysis with original data is performed, the major task is to obtain data from all project managers in a compatible way. Our experience shows that this is possible in principle but time consuming as data may not be available on modern electronic devices and often adaptations between database systems are required.

Step 5. Descriptive analysis: A first step in summarizing the results should be an extensive description of the single papers, including tabulation of relevant elements of each study, such as sample size, data collection procedures, confounder variables, means of statistical analysis, study design, publication year, performing year, geographical setting, etc. This request is also included in the guidelines for publications of meta-analysis that were published by Stroup et al. (2000).

Step 6. Statistical analysis: This includes the analysis of the heterogeneity of the study-specific effects, the calculation of a pooled estimate and the confidence interval as well as a sensitivity analysis. Details are given in the next section on statistical methods.

Step 7. Interpretation of the results: The importance of the sources and magnitude of different biases should be taken into account when interpreting the results. Combining several studies will often give small confidence intervals and suggest a false precision (Egger et al. 1998), but estimates may be biased. For clinical studies, Thompson (1994) has pointed out that the investigation of the heterogeneity between studies will generally give more insight than inspecting the confidence interval of the pooled estimate. This is even more true for a meta-analysis from epidemiological studies. Additionally, the possible effects of publication bias (see below) need to be considered carefully (Copas and Shi 2001).

Step 8. Publication: Guidelines for reporting meta-analyses of observational studies have been published by Stroup et al. (2000). More recently, a large group of methodologists and clinicians revised an existing guidance and checklist focused on the reporting of meta-analyses of randomized controlled trials named PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses, Moher et al. 2009). These guidelines are quite useful for preparing the publication and are also supported by most editors of major medical journals. Especially the detailed description of methods is required so that the analysis could be replicated by others.

36.5 Statistical Analysis

The statistical analysis of aggregated data from published studies was first developed in the fields of psychology and education (Glass 1977; Smith and Glass 1977). These methods have been adopted since the mid-1980s in medicine primarily for randomized clinical trials and are also used for epidemiological data. A recent overview may be found in the review by Sutton and Higgins (2008). We will give a brief outline of some issues of the analysis using an example based on a meta-analysis performed by Sillero-Arenas et al. (1992). This study was one of the first meta-analyses which tried to summarize quantitatively the association between hormone replacement therapy (HRT) and breast cancer in women. Sillero-Arenas et al. based their meta-analysis on 23 case-control and 13 cohort studies. The data extracted from their paper are given in Appendix A of this chapter.

The statistical analysis of MAP is more complex and not covered here.

36.5.1 Single Study Results

A first step of the statistical analysis is the description of the characteristics and the results of each study. Tabulations and simple graphical methods should be employed to visualize the results of the single studies. Plotting the odds ratios and their confidence intervals (so-called forest plot) is a simple way to spot obvious differences between the study results.

Figure 36.1 shows a forest plot of 36 studies investigating the association of HRT and breast cancer in women. Obviously there is a high variability of effects between studies present. Later we will describe how to account for heterogeneity of studies quantitatively.

36.5.2 Publication Bias

An important problem of meta-analysis is publication bias. This bias has received a lot of attention particularly in the area of clinical trials. Publication bias occurs when studies that have non-significant or negative results are published less frequently than positive studies. For randomized clinical trials, it has been shown that even with a computer-aided literature search not all of the relevant studies will be identified (Dickersin et al. 1994). In epidemiological studies, additional problems exist, because often a large number of variables will be collected as potential confounders (Blettner et al. 1999), but only significant or important results are mentioned in the abstract. Therefore, the respective publication may be overlooked for variables which are not mentioned in the abstract. Results of important variables may be published in additional papers, which have often not been planned in advance. In general, publication bias yields a non-negligible overestimation of the risk estimate. Consequently, publication bias should be investigated prior to further statistical analyses.

A simple graphical tool to detect publication bias is the so-called funnel plot. The basic idea is that studies which do not show an effect and which are not statistically significant are less likely to be published. If the sample size or alternatively the precision (i.e., the inverse of the variance) is plotted against the effect, a hole in lower left quadrant is expected. Figure 36.2 shows examples of funnel plots. The left

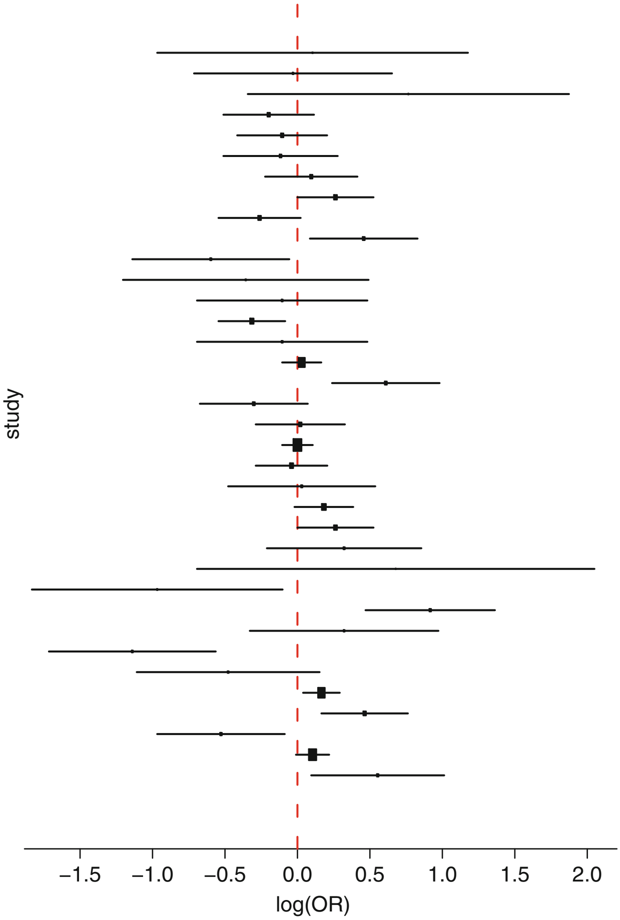


Fig. 36.1 Confidence interval plot of the breast cancer data

subplot of Fig. 36.2 shows a funnel plot with no indication of publication bias. The right subplot shows a so-called apparent hole in the lower left corner. In the case of the right subplot of Fig. 36.2, the presence of publication bias would be assumed.

Figure 36.3 shows a funnel plot for the breast cancer data. No apparent hole in the lower left corner is present. Thus, based on this figure, no publication bias would be assumed.

For a quantitative investigation of publication bias, several methods are available. This may be based on statistical tests; see, for example, Begg and Mazumdar (1994) or Schwarzer et al. (2002). A recent simulation study performed by Macaskill et al. (2001) favored the use of regression methods. The basic idea is to regress the estimated effect sizes $\hat{\theta}_i$ directly on the sample size or the inverse variance σ_i^{-2} as predictor. An alternative idea is to use Egger's regression test (Egger et al. 1997)

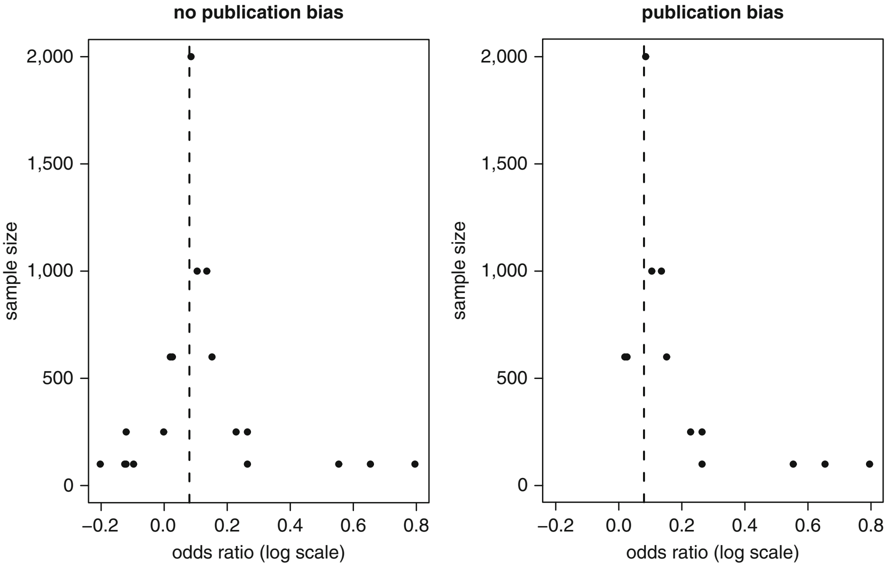


Fig. 36.2 Examples of funnel plots based on simulated data with (*right figure*) and without publication bias present (*left figure*). The *dotted line* shows the true effect

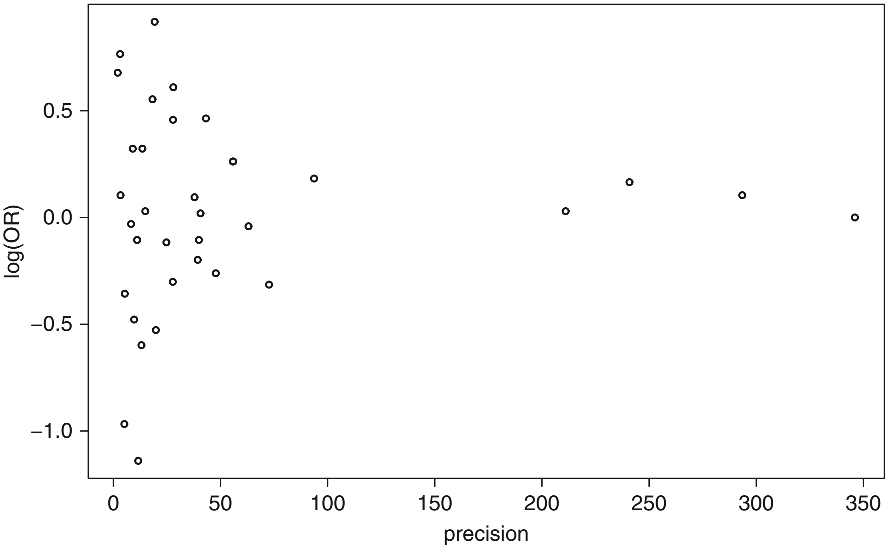


Fig. 36.3 Funnel plot of the breast cancer data

which uses standardized study-specific effects as dependent variable and the corresponding precision as an independent variable. Simulation studies by Macaskill et al. (2001) and by Peters et al. (2006) indicate that the method proposed by Macaskill is superior to Egger's method. Thus, our analysis is restricted to the method by Macaskill which leads to the following regression model

$$\hat{\theta}_i = \alpha + \beta \frac{1}{\sigma_i^2} + \epsilon_i, \quad i = 1, \ldots, k, \quad \epsilon_i \sim N(0, \sigma_i^2).$$

Here, the number of studies to be pooled is denoted by k. In this setting, it is assumed that the estimated treatment effects are independently normally distributed. With no publication bias present, the regression line should be parallel to the x-axis, that is, the slope should be zero. A non-zero slope would suggest an association between sample size or inverse variance, possibly due to publication bias. The estimated regression line in Fig. 36.4 shows no apparent slope. Likewise, the model output (not shown) does not indicate the presence of publication bias for the data at hand.

36.5.3 Estimation of a Summary Effect

Frequently, one of the aims of a meta-analysis is to provide an estimate of the overall effect of all studies combined. Methods for pooling depend on the data available. In general, a two-step procedure has to be applied. First, the risk estimates and

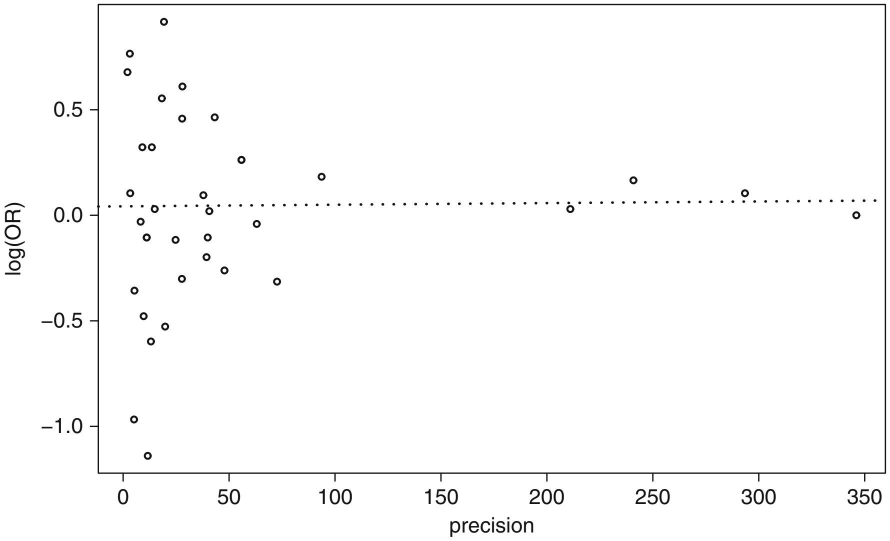


Fig. 36.4 Funnel regression plot of the breast cancer data

variances from each study have to be abstracted from publications or calculated if data are available. Then, a combined estimate is obtained as a (variance based) weighted average of the individual estimates. The methods for pooling based on the 2 × 2 table include the approaches by Mantel-Haenszel and Peto (see Pettiti 1994 for details). If data are not available in a 2 × 2 table, but as an estimate from a more complex model (such as an adjusted relative risk estimate), the Woolf approach can be adopted using the estimates and their (published or calculated) variance resulting from the regression model. This results in a weighted average of the log odds ratios $\hat{\theta}_i$ of the individual studies where the weights w_i are given by the inverse of the study-specific variance estimates $\hat{\sigma}_i^2$. For a discussion of risk measures, see chapter ▶Rates, Risks, Measures of Association and Impact of this handbook. Please note that the study-specific variance is assumed to be fixed and known, although they are based on estimates of the study-specific variances. As a result, the uncertainty associated with the estimation of σ_i^2 is ignored. Thus, in the following, the σ_i are treated as constants and the "hat" notation is omitted. The estimate of the summary effect of all studies is then given by

$$\begin{aligned} \hat{\theta} &= \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}, \\ w_i &= \frac{1}{\sigma_i^2}. \end{aligned}$$

The variance is given by

$$\operatorname{var}(\hat{\theta}) = \frac{1}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}}.$$

Applying this approach to the HRT data leads to a pooled risk estimate of 0.05598 with an estimated variance equal to 0.00051. Transforming this back to the original scale leads to an odds ratio of 1.058 with a 95% confidence interval of (1.012, 1.11). Thus, we would conclude combining all studies that there is a small harmful effect of hormone replacement therapy.

The major assumption here is that of a fixed model, that is, it is assumed that the underlying true exposure effect in each study is the same. The overall variation and, therefore, the confidence intervals will reflect only the random variation within each study but not any potential heterogeneity between the studies.

Figure 36.5 displays this idea. Whether pooling of the data is appropriate should be decided after investigating the heterogeneity of the study results. If the results vary substantially, no pooled estimator should be presented or only estimators for selected subgroups should be calculated (e.g., combining results from case-control studies only).

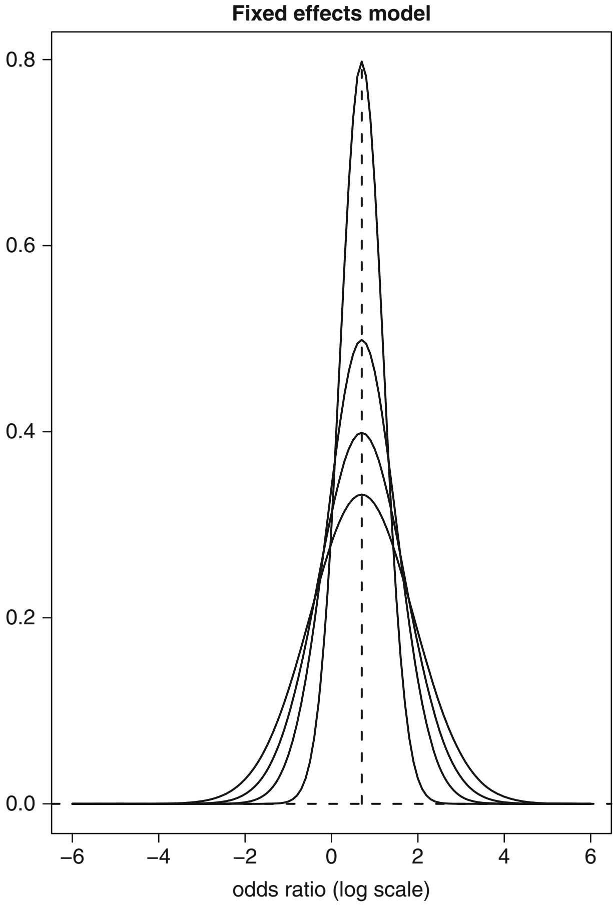


Fig. 36.5 Fixed effects model: common effect with different study variances

36.5.4 Heterogeneity

The investigation of heterogeneity between the different studies is a main task in each review or meta-analysis (Thompson 1994). For the quantitative assessment of heterogeneity, several statistical tests are available (Pettiti 1994; Paul and Donner 1989). A simple test for heterogeneity is based on the following test statistic:

$$\chi^2_{\text{het}} = \sum_{i=1}^{k} \frac{(\hat{\theta} - \hat{\theta}_i)^2}{\sigma_i^2} \sim \chi^2_{k-1} \qquad (36.1)$$

which under the null hypothesis of heterogeneity follows a χ^2 distribution with $k-1$ degrees of freedom. Hence, the null hypothesis is rejected if χ^2_{het} exceeds the $1-\alpha$

quantile of χ^2_{k-1} denoted as $\chi^2_{k-1,1-\alpha}$. For the data at hand, we clearly conclude that there is heterogeneity present ($\chi^2_{\text{het}} = 116.076$, df $=35$, p-value: 0.00000). Thus, using a combined estimate is at least questionable. Pooling the individual studies and performing this test can be done with any statistical package capable of weighted least squares regression. In part B of the Appendix of this chapter shows a SAS program which provides the results obtained so far. A major limitation of formal heterogeneity tests like the one presented before is, however, their low statistical power to detect any heterogeneity present.

A more powerful method is given by model-based approaches. A model-based approach has the advantage that it can be used to test specific alternatives and thus has a higher power to detect heterogeneity. So far, we considered the following simple fixed effects model:

$$\theta_i = \theta + \varepsilon_i, \quad i = 1, \ldots, k, \quad \varepsilon_i \sim N(0, \sigma_i^2).$$

Obviously, this model is not able to account for any heterogeneity, since deviations from θ_i and θ are assumed to be explained only by random error.

Thus, alternatively a random effects model should be considered. This model incorporates variation between studies. It is assumed that each study has its own (true) exposure effect and that there is a random distribution of these true exposure effects around a central effect. This idea is presented in Fig. 36.6. Frequently, it is assumed that the individual study effects follow a normal distribution with mean θ_i and variance σ_i^2 and the random distribution of the true effects is again a normal distribution with variance τ^2. In other words, the random effects model allows non-homogeneity between the effects of different studies. This leads to the following model:

$$\theta_i = \theta + b_i + \epsilon_i, \quad i = 1, \ldots, k, \quad b_i \sim N(0, \tau^2), \quad \epsilon_i \sim N(0, \sigma_i^2). \tag{36.2}$$

The observed effects from the different studies are used to estimate the parameters describing the fixed and random effects. This may be done using maximum-likelihood procedures. In the following sections, two further methods of estimating the heterogeneity variance in a random effects model are presented.

36.5.4.1 The Heterogeneity Variance Estimator Proposed by DerSimonian and Laird

The widely used approach by DerSimonian and Laird (1986) applies a method of moments to obtain an estimate of τ^2. Taking the expectation of (36.2) leads to $E(\theta_i) = \theta$ and calculating the variance leads to $\text{var}(\theta_i) = \text{var}(b_i) + \text{var}(\varepsilon_i) = \tau^2 + \sigma_i^2 = \sigma_i^{*^2}$ assuming that b_i and ϵ_i are independent. The heterogeneity variance τ^2 is unknown and has to be estimated from the data. The method by DerSimonian and Laird equates the heterogeneity test statistic (36.1) to its expected value. This expectation is calculated under the assumption of a random effects model and given

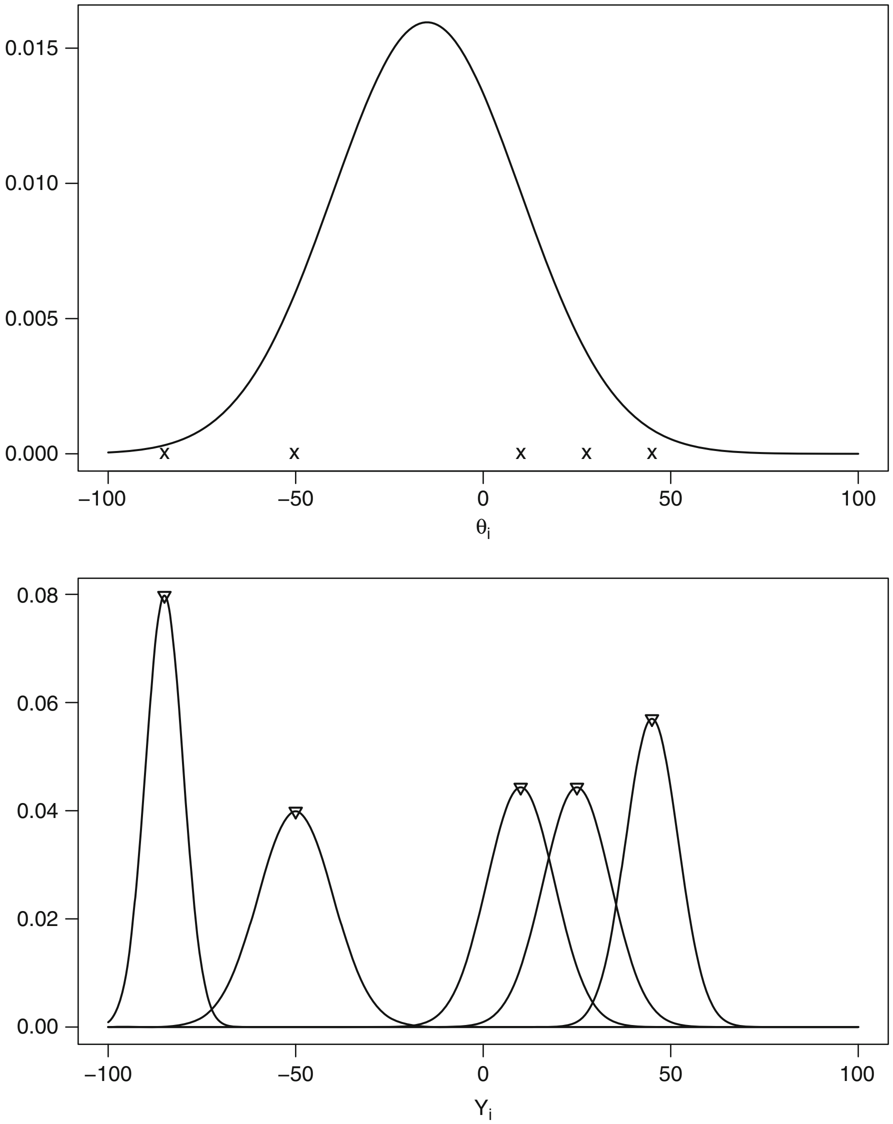


Fig. 36.6 Random effects model: variable effects drawn from a population of study effects

by $E(\chi^2_{\text{het}}) = k-1+\tau^2\left(\sum w_i - \frac{\sum w_i^2}{\sum w_i}\right)$. The weights w_i are those defined in (36.1). Equating χ^2_{het} to its expectation and solving for τ^2 gives

$$\hat{\tau}^2 = [\chi^2_{\text{het}} - (k-1)] / \left(\sum w_i - \frac{\sum w_i^2}{\sum w_i}\right).$$

In case $\chi^2_{\text{het}} < k-1$, the estimator $\hat{\tau}^2$ is truncated to zero. Thus, the pooled estimator $\hat{\theta}_{\text{DL}}$ under heterogeneity can be obtained as weighted average:

$$\hat{\theta}_{\text{DL}} = \frac{\sum_{i=1}^{k} w_i^* \hat{\theta}_i}{\sum_{i=1}^{k} w_i^*} \tag{36.3}$$

$$\text{with} \qquad w_i^* = \frac{1}{\sigma_i^{*2}} = \frac{1}{\hat{\tau}^2 + \sigma_i^2} \quad \text{we obtain} \tag{36.4}$$

$$\hat{\theta}_{\text{DL}} = \frac{\sum_{i=1}^{k} \hat{\theta}_i / (\hat{\tau}^2 + \sigma_i^2)}{\sum_{i=1}^{k} 1/(\hat{\tau}^2 + \sigma_i^2)}.$$

The variance of this estimator is given by:

$$\begin{aligned} \text{var}(\hat{\theta}_{\text{DL}}) &= \frac{1}{\sum_{i=1}^{k} \frac{1}{\sigma_i^{*2}}} \\ &= \frac{1}{\sum_{i=1}^{k} \frac{1}{\hat{\tau}^2 + \sigma_i^2}}. \end{aligned}$$

The between-study variance τ^2 can also be interpreted as a measure for the heterogeneity between studies. For our example, we obtain a pooled DerSimonian-Laird estimate of 0.0337 with heterogeneity variance equal to 0.0453. The variance of the pooled estimator is given by 0.0024. Transformed back to the original scale, we obtain an odds ratio of $OR = 1.034$ with 95% CI (0.939, 1.139). Based on this analysis, we would conclude that after adjusting for heterogeneity, this meta-analysis does not provide evidence for an association between HRT replacement therapy and breast cancer in women. It should be noted that within this approach, the study-specific variances are assumed to be known constants. This can lead to a considerable bias when pooling estimates using the DerSimonian-Laird estimator as demonstrated by Böhning et al. (2002).

36.5.4.2 Another Estimator of τ^2: The Simple Heterogeneity Variance Estimator

Sidik and Jonkman (2005) proposed a simple method for the estimation of the heterogeneity variance τ^2. As before, they consider a model as in (36.2) with $b_i \sim N(0, \tau^2)$. Now, for the purpose of their estimation method random effects model, they reparameterize the variance as

$$\text{var}(\hat{\theta}_i) = \tau^2 + \sigma_i^2 = \tau^2 \frac{\sigma_i^2 + \tau^2}{\tau^2} = \tau^2 (r_i + 1)$$

with $r_i = \sigma_i^2/\tau^2$. Then, the problem of estimating τ^2 is cast into the framework of simple linear regression:

$$E(\hat{\theta}) = X\theta$$
$$\text{var}(\hat{\theta}_1) = \tau^2 V.$$

Here, X is a vector of 1s with dimension $k \times 1$, and V is a diagonal matrix with elements $r_1 + 1, \ldots, r_k + 1$. In the framework of the usual weighted least squares, an estimator of θ is obtained as

$$\hat{\theta}_\nu = \frac{\sum_{i=1}^k \nu_i^{-1}\hat{\theta}_i}{\sum_{i=1}^k \nu_i^{-1}}$$

with $\nu_i = r_i + 1$. Please note that this is equivalent to (36.3) and (36.4). The advantage of casting the problem into the usual weighted least squares approach is that an estimator of τ^2 can be obtained as the usual weighted residual sum of squares as follows:

$$\hat{\tau}^2 = \frac{1}{k-1}\sum_{i=1}^{k} \nu_i^{-1}(\hat{\theta}_i - \hat{\theta}_\nu)^2. \tag{36.5}$$

Of course, an estimator of τ is needed to compute the ratios r_i of the within-study variance σ_i^2 and the between-study variance τ^2. Here, Sidik and Jonkman propose to use the empirical variance of the study-specific estimates $\hat{\theta}_i$

$$\hat{\tau}_0^2 = \frac{1}{k}\sum_{i=1}^{k}(\hat{\theta}_i - \bar{\theta})^2.$$

Plugging this into (36.5) leads to

$$\hat{\tau}_{SH}^2 = \frac{1}{k-1}\sum_{i=1}^{k} \hat{\nu}^{-1}(\hat{\theta}_i - \hat{\theta}_\nu)^2$$

with $\hat{\nu}_i = \hat{r}_i + 1$ and $\hat{r}_i = \sigma_i^2/\hat{\tau}_0^2$. Please note that this estimator is strictly positive in contrast to the DerSimonian-Laird estimator. This estimator will be referred to as the heterogeneity variance estimator SH. For the data at hand, an estimate of τ^2 equals 0.199.

36.5.4.3 Heterogeneity Variance Estimation by Likelihood-Based Methods

Besides the moment-based method by DerSimonian and Laird or the simple heterogeneity variance estimator, estimates of τ^2 can be obtained using likelihood-based methods. See, for example, the tutorials by Normand (1999) and van Houwelingen et al. (2002) for more details. The Appendix B and C of this chapter gives a SAS code to estimate the fixed and random effects models based on likelihood methods with the SAS program *proc mixed*. Estimates based on likelihood methods offer the advantage that they provide the option to formally test which model is appropriate for the data by applying the likelihood ratio test or penalized criteria such as the Bayesian information criterion (BIC). The BIC is obtained by the formula $\text{BIC} = -2 \times \log \text{Likelihood} + \log(k) \times q$ where q is the number of parameters in the model and k denotes the number of studies.

When using random effects models, another topic of interest is the form of the random effects' distribution. Besides a parametric distribution for the random effects, a discrete distribution may be assumed. Here, we suppose that the study-specific estimators $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ are coming from q subpopulations $\theta_j, j = 1, \ldots, q$. Again, assuming that the effect of each individual study follows a normal distribution

$$f(\hat{\theta}_i, \theta_j, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(\hat{\theta}_i - \theta_j)^2}{2\sigma_i^2}}, \quad j = 1, \ldots, q,$$

we obtain a finite mixture model

$$f(\hat{\theta}_i, P) = \sum_{j=1}^{q} f(\hat{\theta}_i, \theta_j, \sigma_i^2) p_j .$$

The parameters of the distribution P

$$P \equiv \begin{bmatrix} \theta_1 & \ldots & \theta_q \\ p_1 & \ldots & p_q \end{bmatrix} \quad \text{with} \quad p_j \geq 0, \quad j = 1, \ldots, q,$$
$$p_1 + \cdots + p_q = 1,$$

need to be estimated from the data. The mixing weights p_j denote the a priori probability of an observation of belonging to a certain subpopulation with parameter θ_j. Please note that also the number of components q needs to be estimated as well. Estimation may be done with the program C.A.MAN (Schlattmann and Böhning 1993; Böhning et al. 1998) or more recently with the R package CAMAN (Schlattmann 2009). For the HRT data, we find a solution with three components which gives an acceptable fit to the data

```
weight:      0.2804 parameter:     -0.3365
weight:      0.5671 parameter:      0.0778
weight:      0.1524 parameter:      0.5446

Log-Likelihood at iterate :       -17.6306
```

Table 36.1 Model comparison for the breast cancer data

	Residual	Estimates (SE)			
Method	Hetero	Intercept	Het. ($\hat{\tau}^2$)	log Lik.	BIC
Fixed	None	0.056 (0.023)	–	−33.19	70.0
Mixed	Additive	0.027 (0.061)	0.086	−18.65	44.4
FM	Additive		0.079	−17.63	53.2

Here, the weights correspond to the mixing weights p_j, and the parameter corresponds to the subpopulation mean θ_j. These results imply that about 28% of the studies show a protective effect of HRT, whereas the majority of the studies show a harmful effect. About 57% of the studies show an increased log(risk) of 0.08, and 15% of the studies show a log(odds ratio) of 0.54. Thus, using a finite mixture model (FM), we find again considerable heterogeneity where the majority of studies find a harmful effect of hormone replacement therapy. It is noteworthy that a proportion of studies find a beneficial effect. Of course, this needs to be investigated further. One way to do this would be to classify the individual studies using the finite mixture model. Doing so, we find that, for example, study nine from the data given in the Appendix A belongs to this category. This is a case-control study for which no information about confounder adjustment is available. This would be a starting point for a sensitivity analysis. Table 36.1 gives an overview about the models fitted so far. These include the fixed effects model with a BIC value of 70.0, the mixed effects model using a normal distribution for the random effects with a BIC value of 44.4. The finite mixture model (FM) has a BIC value of 53.2. Thus, based on Table 36.1, it is quite obvious that a fixed effects model does not fit the data very well and that a random effects model should be used. Of course, the question remains which random effects model to choose for the analysis. Based on the BIC criterion given in Table 36.1, one would choose the parametric mixture model provided the assumption of a normal distribution of the random effects is justifiable. This can be investigated, for example, by a normal quantile-quantile plot of the estimated individual random effects given by the parametric model. For the data at hand, the assumption of normally distributed random effects appears reasonable; thus, we would choose the parametric mixture model.

36.5.4.4 Further Aspects of Heterogeneity

It should be noted that in general random effects models yield larger variance and confidence intervals than fixed effects models because a between-study component τ^2 is added to the variance. If the heterogeneity between the studies is large, τ^2 will dominate the weights and all studies will be weighted more equally (in the random effects model weight decreases for larger studies compared to the fixed effects model).

Furthermore, pooling in the presence of heterogeneity may be seriously misleading. Heterogeneity between studies warrants careful investigation of the sources of the differences. If there are a sufficient number of different studies available, further analyses, such as “meta-regression,” may be used to examine the sources of heterogeneity (Greenland 1987, 1994). One specific problem that occurs when binary endpoints are considered is that different risk measurements are used. While in case-control studies, an odds ratio is estimated, cohort studies yield an estimate for the relative risk. Although for studies in which rare events are investigated, odds ratios and relative risk estimates are very similar, this is not the case in studies where diseases with a higher prevalence are investigated. Odds ratio and relative risk differ and should not be combined. The problem that arises if studies with different designs are combined has not been well studied. We believe that no pooled estimate should be calculated combining data from case-control and cohort studies. There are many sources of heterogeneity, not only the size of risk estimates.

36.5.4.5 Meta-Regression

An important method for investigating heterogeneity is sensitivity analysis, for example, to calculate pooled estimators only for subgroups of studies (according to study type, quality of the study, period of publication, etc.) to investigate variations of the odds ratio. An extension of this approach is meta-regression as proposed by Greenland (1987); see also Thompson and Sharp (1999). The principal idea of meta-regression is once heterogeneity is detected to identify sources of heterogeneity by inclusion of known covariates.

For the breast cancer meta-analysis example a potential covariate is study type, case-control studies may show different results than cohort studies due to different exposure assessment. For our data, case-control studies are coded as $x_{i1} = 0$, and cohort studies are coded as $x_{i1} = 1$.

The fixed effect model is now

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i, \quad \varepsilon_i \sim N(0, \sigma_i^2), \quad i = 1, \ldots, k.$$

Here, we find that cohort studies identify an association between HRT and breast cancer based on the regression equation $\hat{\theta}_i = 0.0015 + 0.145$ for a cohort study. Obviously, cohort studies come to results different form case-control studies. Clearly, after adjustment for covariates, the question remains if there is still residual heterogeneity present. Again, we can analyze the data using a random effects model in this case with a random intercept:

$$\theta_i = \beta_0 + \beta_1 x_{i1} + b_i + \epsilon_i, \quad b_i \sim N(0, \tau^2), \quad \varepsilon_i \sim N(0, \sigma_i^2).$$

For this model, the regression equation for the fixed effects gives now for a cohort study $\hat{\theta}_i = -0.009 + 0.1080$, and the corresponding heterogeneity variance is estimated as $\hat{\tau}^2 = 0.079$.

Table 36.2 compares fixed and random effects models for the HRT data. The table shows models with and without an estimate for the slope. Model selection can be based again on the BIC criterion. Apparently based on the BIC criterion, both fixed

Table 36.2 Comparison of fixed and random effects models

	Residual	Estimates (SE)				
Method	Hetero	Intercept	Slope	Het. ($\hat{\tau}^2$)	−log Lik.	BIC
Fixed	None	0.056 (0.023)	–	–	−33.85	70.0
Fixed	None	0.0014 (0.029)	0.145 (0.046)	–	−28.36	63.9
Mixed	Additive	0.027 (0.061)	–	0.086	−18.65	44.4
Mixed	Additive	−0.009 (0.072)	0.108 (0.126)	0.079	−18.25	47.3

effects models do not fit the data very well since their BIC values are considerably higher than those of the random effects models. Please note that if only the fixed effects models would be considered, this meta-analysis would show that cohort studies show a harmful effect. Comparing the mixed effects models in Table 36.2, the model with the covariate does not provide an improved fit of the data. The log-likelihood is only slightly larger, and penalizing the number of parameters leads to a larger BIC value for the mixed effect model with the covariate. Another interesting point is to compare the heterogeneity variance estimated by both models. Here, there is no substantial portion of heterogeneity explained by the covariate, since the heterogeneity variance is reduced to 0.079 from 0.086. From a statistical point of view, further covariates need to be identified and included into the model. From a public health point of view, the conclusion is perhaps less straightforward. Although inclusion of the covariate study type does not explain the heterogeneity of the studies very well, we find that cohort studies find a harmful effect. One might argue that although these results are far from perfect, they should not be ignored as absence of evidence does not imply evidence of absence. Looking back at these data in the light of the results from the woman health initiative (WHI) study (Rossouw et al. 2002), it becomes clear that caution is required in the analysis and interpretation of meta-analyses of observational studies. The major finding of the WHI study was that the group of subjects undergoing treatment with combined HRT in the form of Prempro (0.625 mg/day conjugated equine estrogens (CEE) +2.5 mg/day medroxyprogesterone acetate) was found to have increased risk of breast cancer (hazard ratio = 1.26, 95% CI: 1.00–1.59) and no apparent cardiac benefit. This is contradictory to the prior belief that HRT provides cardiovascular benefit. As a result, although several benefits were considered, these interim findings at 5 years were deemed sufficiently troubling to stop this arm of the trial at 5.2 years.

36.6 Interpretation of the Results of Meta-Analysis of Observational Studies

The example from above shows that the interpretation of the results of a meta-analysis should not only discuss the pooled estimator and the confidence interval but should focus on the examination of the heterogeneity between the results of the studies. Strength and weaknesses as well as potential bias should be discussed.

36.6.1 Bias

For epidemiological studies in general, the main problem is not the lack of precision and the random error but the fact that results may be distorted by different sources of bias or confounding; for a general overview of the problem of bias, see Hill and Kleinbaum (2000). That means that the standard error (or the size of the study) may not be the best indicator for the weight of a study. If more or better data are collected on a smaller amount of subjects, results may be more accurate than in a large study with insufficient information on the risk factors or on confounders. The assessment of bias in individual studies is therefore crucial for the overall interpretation.

The central problem of meta-analyses of clinical trials is publication bias that has already been a topic in a paper by Berlin et al. as early as 1989 and is still a topic of recent methodological investigations (see, e.g., Copas and Shi 2001). This bias has received a lot of attention particularly in the area of clinical trials. As mentioned in Sect. 36.5.2, publication bias yields in general a non-negligible overestimation of the risk estimate. However, as Morris (1994) has pointed out, there exist little systematic investigations of the magnitude of the problem for epidemiological studies. A major worry is that non-significant results are neither mentioned in the title nor in the abstract and publications and may be lost in the retrieval process.

36.6.2 Confounding

Another problem arises because different studies adjust for different confounding factors. It is well known that the estimated effect of a factor of interest is (strongly) influenced by the inclusion or exclusion of other factors, in the statistical model if these factors have an influence on the outcome and if they are correlated with the risk factor of interest. Combining estimates from several studies with different ways of adjusting for confounders yields biased results. Using literature data only, crude estimates may be available for some of the studies, model-based estimates for others. However, as the adjustment for confounders is an important issue for the assessment of an effect in each single study, it is obvious that combining these different estimates in a meta-analysis may not give meaningful results. It is necessary to use "similar" confounders in each study to adjust the estimated effect of interest in the single studies. In general, that would require a reanalysis of the single studies. Obviously, that requires the original data and a MAP is needed for this purpose.

36.6.3 Heterogeneity

In epidemiological research, different study designs are in use, and none of them can be considered as a gold standard as the randomized clinical trial for therapy studies. Therefore, it is necessary to evaluate the comparability of the single designs before summarizing the results. Often, case-control studies, cohort studies, and cross-sectional studies are used to investigate the same questions, and results of those studies need to be combined. Egger et al. (2001b) pointed out several examples in which results from case-control studies differ from those of cohort studies. For example, in a paper by Boyd et al. (1993), it was noted that cohort studies show no association between breast cancer and saturated fat intake, while the same meta-analysis using results from case-control studies only revealed an increased, statistically significant risk. Other reasons for heterogeneity may be different uses of data collection methods, different control selection (e.g., hospital vs. population controls), and differences in case ascertaining. Differences could be explored in a formal sensitivity analysis but also by graphical methods (funnel plot). However, meta-analyses from published data provide only limited information if the reasons for heterogeneity shall be investigated in depth.

The problem of heterogeneity can be well demonstrated with nearly any example of published meta-analysis. For example, Ursin et al. (1995) investigated the influence of the body mass index (BMI) on the development of premenopausal breast cancer. They included 23 studies of which 19 are case-control studies and 4 are cohort studies. Some of these studies were designed to investigate BMI as risk factor, others measured BMI as confounders in studies investigating other risk factors. It can only be speculated that the number of unpublished studies in which BMI was mainly considered as a confounder and did not show a strong influence on premenopausal breast cancer is non-negligible and that this issue may result in some bias. As is usual practice in epidemiological studies, relative risks were provided for several categories of BMI. To overcome this problem, the authors estimated a regression coefficient for the relative risk as a function of the BMI; however, several critical assumptions are necessary for this type of approach. The authors found severe heterogeneity across all studies combined (the p-value of a corresponding test was almost zero). An influence of the type of study (cohort study or case-control study) was apparent. Therefore, no overall summary is presented for case-control and cohort studies combined. One reason for the heterogeneity may be the variation in adjustment for confounders. Adjustment for confounders other than age was used only in 10 out of the 23 studies.

36.7 Network Meta-Analysis

Network meta-analyses are a recent development in meta-analyses, especially for randomized clinical trials, and are also referred to as multiple treatments meta-analyses. They are applied when the difference between two or more medical treatments should be investigated, but no or only few head-to-head comparison studies are available. In this situation, studies comparing the treatments of interest with a common comparator can be used for relative effect inferences. For example, there are randomized studies comparing treatment B versus standard treatment

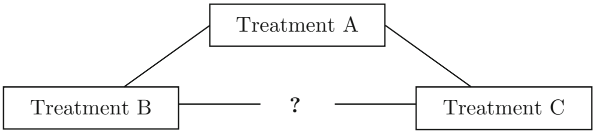


Fig. 36.7 Indirect comparison between treatment B and treatment C

A and furthermore randomized studies comparing treatment C versus treatment A (shown graphically in Fig. 36.7). A previously unresearched effect of treatment B versus treatment C can then be evaluated by an indirect comparison using the available studies.

36.7.1 Statistical Methods

In the setting of Fig. 36.7, let $\hat{\theta}_{BA}$ and $\hat{\theta}_{CA}$ be the estimated summary effects of the corresponding studies. These effects, for example, log odds ratios for binary outcome or differences in means for continuous data, can be combined in a network meta-analysis as follows:

$$\hat{\theta}_{BC} = \hat{\theta}_{BA} - \hat{\theta}_{CA}.$$

Extending this network, different paths for the same treatment comparison can also exist.

If further randomized studies were available comparing treatment D with treatments B and C, respectively (see Fig. 36.8), the effect θ_{BC} could be estimated on a second way:

$$\hat{\theta}_{BC} = \hat{\theta}_{BD} - \hat{\theta}_{CD}.$$

The network of different treatment comparisons can be arbitrarily extended to use all available evidence. Accordingly, several treatments can be compared simultaneously based on pairwise comparative or multiarm studies. If direct and indirect comparisons are included in a network meta-analysis, it is also named as mixed treatment comparisons (MTC) meta-analysis (Lu and Ades 2004). For example, Elliott and Meyer (2007a) investigated the effect of five antihypertensive drugs and placebo on incident diabetes mellitus based on 22 clinical trials with comparisons between two and three treatments. Figure 36.9 shows the network with the summary estimates for pairwise comparisons of traditional meta-analyses and each included study. None of these 22 trials performed a direct comparison between angiotensin-converting enzyme (ACE) inhibitors and angiontensin receptor blockers (ARB).

The effect estimates used for indirect comparisons or the combination of different network paths may be based on fixed or random effects models. In the simplest case, independent estimates of different paths can be synthesized by an average, weighted by the inverse of their respective variances. Furthermore, Bayesian methods and hierarchical models can be used to analyze indirect comparisons as well as more complex data structures such as multiarm trials, especially when

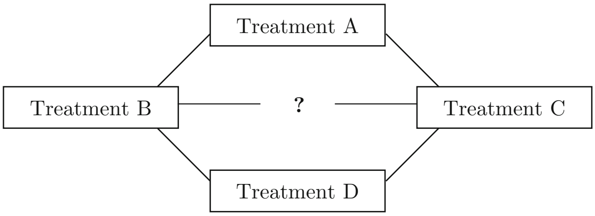


Fig. 36.8 Simple network meta-analysis

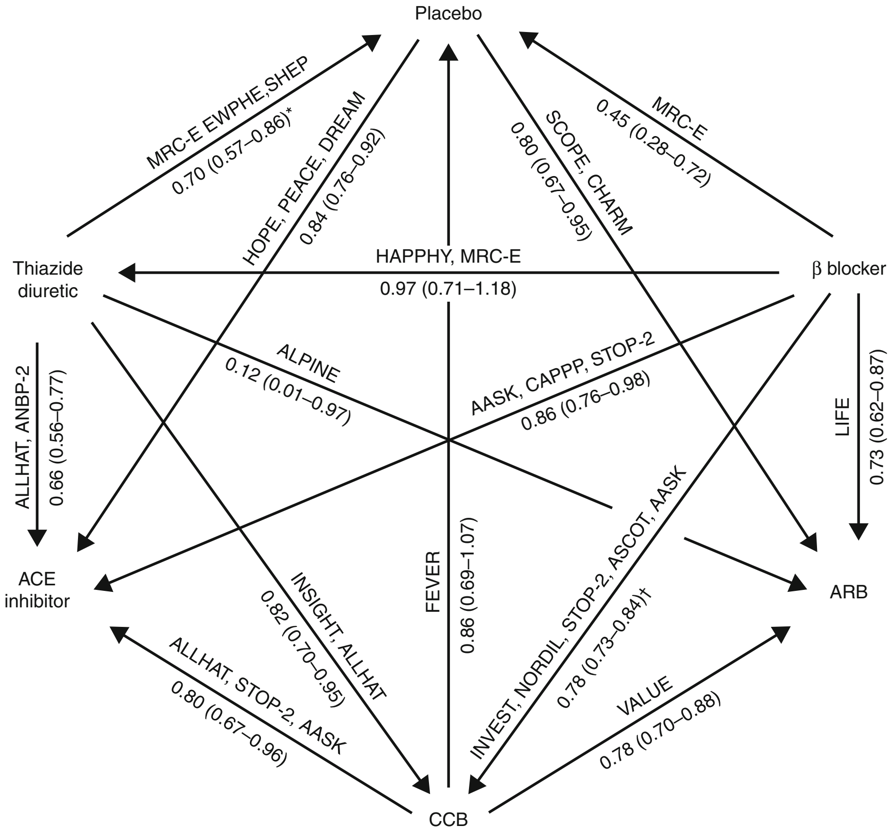


Fig. 36.9 A network of antihypertensive drug clinical trials (Reprinted from Elliott and Meyer 2007b, with permission from Elsevier)

additional information is available (Lu and Ades 2004). Adjustments for covariates and the use of meta-regression methods are also possible in network meta-analyses. Nixon et al. (2007) presented an MTC Bayesian meta-regression model to estimate the efficacy of biological treatments in rheumatoid arthritis adjusted for the both study-level covariates, disease duration and severity of disease.

36.7.2 Limitations

The reasons for potential bias, as already discussed for common meta-analysis, can also lead to distortions in network meta-analysis. But since an indirect comparison requires at least two meta-analyses, there is assumedly more scope for error.

For a valid indirect comparison, the studies should be methodologically and clinically comparable. Also, it is important to synthesize the evidence from different randomized clinical trials by comparing their treatment effects, such as log odds ratios. Considering only single treatment arms of the original controlled studies would break the randomization (Bucher et al. 1997). For example, the observations of treatment arm B and treatment arm C of the original by treatment A controlled studies (see Fig. 36.7) should not be pooled. Although the data originate from randomized controlled trials, they are reduced to the equivalent of those that come from an observational study. The comparison of effectiveness can therefore be biased by several confounders. Estimates of indirect comparisons based only on single treatment arms of randomized controlled trials are sometimes referred to as "unadjusted indirect estimates" and should be avoided.

In a network meta-analysis, not only the heterogeneity in treatment effects between studies of each pairwise contrast should be discussed, but also different network paths for the same comparison. For example, the estimations $\hat{\theta}_{BA} - \hat{\theta}_{CA}$ and $\hat{\theta}_{BD} - \hat{\theta}_{CD}$ in Fig. 36.8 can disagree. Particularly in MTC meta-analyses, inconsistencies in treatment effects obtained from direct and indirect comparisons of the same treatments may exist (Song et al. 2003). Baker and Kramer (2002) highlighted examples where the transitivity assumption between a direct and an indirect comparison is not given, for example, due to differences in patient or trial characteristics. In addition, imperfections in study design or analysis may lead to inconsistencies. For example, the observed treatment effect in open-label or ineffectively blinded studies might depend on whether the treatment is new or used as an active control which may alter the expectation of researchers and patients. This may in turn influence the observed treatment effect (see Lumley 2002). Song et al. (2009) surveyed systematic reviews published between 2000 and 2007 with indirect comparisons and investigated basic assumptions and methodological problems in the application. The assumption of consistency was explicitly mentioned in only 18 of the 30 included reviews where direct and indirect estimates were compared or combined.

The investigation of inconsistency, or also known as incoherence, is possible when there are closed loops such as in the connected polygonal network in Fig. 36.9. Independent estimates of different network paths can be tested, for example, with the simple χ^2-test like in the testing of heterogeneity (see Sect. 36.5.4 and Caldwell et al. 2009). Lumley (2002) offered a quantitative estimation procedure using linear mixed models, in which not only the heterogeneity but also the inconsistency can be fitted as a further random effect. Lu and Ades (2006) set the estimation of inconsistency in the framework of Bayesian hierarchical models. However, large inconsistency needs, as large heterogeneity, a closer look for possible reasons. The adjustment for covariates and the use of meta-regression methods are another

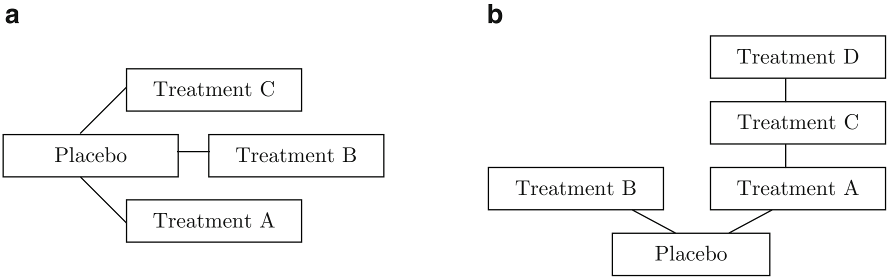


Fig. 36.10 Different network geometries: (**a**) radiating star (**b**) linear structure

possibility to address the inconsistency and also the heterogeneity within comparisons (Cooper et al. 2009). But since inconsistency is a property of loops, it is not easily possible to ascertain which specific contrast is responsible for the inconsistency, and it is an open question to which extent unexplained inconsistency is acceptable in an MTC meta-analysis. For a radiating or linear network meta-analysis structure, the estimation of inconsistency is impossible. Pattern (a) in Fig. 36.10 arises, for example, if there are only placebo-controlled studies and the active treatments are not compared to each other. If new treatments are tested only against the current standard, pattern (b) in Fig. 36.10 might result. Fewer cross-links and longer paths, for example, between treatment B and treatment C in Fig. 36.10, lead to greater uncertainty of the treatment effect estimation. But the information in these networks can be used, for example, in the planning of future trials to directly test the comparison where no direct evidence exists or to enhance the network. Salanti et al. (2008) discussed network geometry and asymmetry in this context, namely, that specific treatments or comparisons are represented more heavily than others.

Comparing the estimation accuracy between direct and indirect comparisons shows that under the assumption of statistical independence of different studies and consistent estimation accuracy, the variance of the effect estimator in indirect comparisons is greater than in direct comparisons. Assuming, for example, that the effect estimates of all studies have the same variance σ^2, the variance of the fixed-effect inverse variance estimator of the direct comparisons based on $2k$ studies would be

$$\text{var}(\hat{\theta}_{BC_{\text{direct}}}) = \sigma^2/2k.$$

However, the variance of the indirect comparison based on k studies for each comparison would be

$$\text{var}(\hat{\theta}_{BC_{\text{indirect}}}) = \text{var}(\hat{\theta}_{BA_{\text{direct}}}) + \text{var}(\hat{\theta}_{CA_{\text{direct}}}) = \sigma^2/k + \sigma^2/k = 2\sigma^2/k.$$

Thus, for the indirect approach, four times as many studies are needed to have the same power as the direct comparisons provide (Glenny et al. 2005).

Consequently, indirect comparisons are not preferable to direct comparisons, but they are useful in planning direct comparisons and where head-to-head comparisons are not possible. Both can be combined to include all the existing evidence and to obtain more precise estimates when direct comparisons are not sufficiently available (Higgins and Whitehead 1996). Song et al. (2008) hypothesized as to whether adjusted indirect comparisons may provide less biased results than direct comparisons under certain circumstances. For example, when an indirect comparison is based on studies with the same extent of bias, adjusted indirect comparison could counterbalance this. However, discrepancies between both approaches should be carefully investigated for potential explanations whenever possible.

36.8 Conclusions

Despite the many problems, there is an immense need to summarize current knowledge, for example, to assess the consequence of human exposure to environmental exposure. For this task, all available data and information will be needed, and meta-analysis is becoming increasingly influential. Particularly where the previously conducted epidemiological studies have provided inconsistent results, a meta-analysis may give some insight. As discussed, a major impediment for meta-analysis of epidemiological data is the heterogeneity across studies in their design, data collection methods, and analyses performed. The statistical combination of risk estimates should not be the central component of a meta-analysis using published data. An expert group in cooperation with the US Environmental Protection Agency was recently established to discuss the use of meta-analyses in environmental health studies. One of the objectives of this group was also to develop a consensus on "when meta-analysis should or should not be used" (Blair et al. 1995). There is always a danger that meta-analysis of observational studies produces precise looking estimates which are severely biased. This should be kept in mind as more and more public health regulators and decision-makers may rely on the results of a meta-analysis.

Appendices

A. Data

The listing shows the effect measure on the log-scale, the corresponding variance and the study type of each of the 36 studies analyzed in the meta-analysis by Sillero-Arenas et al. (1992):

```
data sillar;
input study or est type;
cards;
1   0.10436  0.299111  0
2  -0.03046  0.121392  0
3   0.76547  0.319547  0
```

```
4   -0.19845  0.025400  0
5   -0.10536  0.025041  0
6   -0.11653  0.040469  0
7    0.09531  0.026399  0
8    0.26236  0.017918  0
9   -0.26136  0.020901  0
10   0.45742  0.035877  0
11  -0.59784  0.076356  0
12  -0.35667  0.186879  0
13  -0.10536  0.089935  0
14  -0.31471  0.013772  0
15  -0.10536  0.089935  0
16   0.02956  0.004738  0
17   0.60977  0.035781  0
18  -0.30111  0.036069  0
19   0.01980  0.024611  0
20   0.00000  0.002890  0
21  -0.04082  0.015863  0
22   0.02956  0.067069  0
23   0.18232  0.010677  0
24   0.26236  0.017918  1
25   0.32208  0.073896  1
26   0.67803  0.489415  1
27  -0.96758  0.194768  1
28   0.91629  0.051846  1
29   0.32208  0.110179  1
30  -1.13943  0.086173  1
31  -0.47804  0.103522  1
32   0.16551  0.004152  1
33   0.46373  0.023150  1
34  -0.52763  0.050384  1
35   0.10436  0.003407  1
36   0.55389  0.054740  1
run;
```

B. Elementary Analysis with SAS

SAS code for the elementary analysis using weighted least squares:

```
/* calculation of weights */
data sillar;
set sillar;
weight =1./est;
run;
```

```
/* intercept only */
proc glm data=sillar;           /* use proc GLM with data set sillar      */
model logor=/solution inverse ; /* Show  solution                         */
                                /* Show inverse of weighted design matrix */
weight weight;                  /*  weights 1./variance                   */
run;
```

This gives the following shortened output:

```
The GLM Procedure
Dependent Variable: logor
Weight: weight
               Sum of
Source     DF  Squares     Mean Square    F Value    Pr > F
Model       1   6.1683128  6.1683128       1.86       0.1813
Error      35 116.0756869  3.3164482
Un.Total   36 122.2439997

Parameter  df   Estimate     SE          t Value   Pr > |t|
Intercept  1    0.0559813731 0.04104847   1.36      0.1813
```

Please note that for performing a meta-analysis the standard error given by the program must be divided by the root mean square error in order to obtain the standard error of the pooled estimate. In order to avoid additional calculations the SAS output giving the inverse of the weighted design matrix gives the desired variance. The test of heterogeneity is given by the residual sum of squares as indicated by formula (36.4). This result can also be obtained using the SAS code for the fixed effect model based on maximum likelihood:

```
proc mixed method=ml data=sillar; /* Use proc mixed (ML
                                     estimation)              */
class study;                      /* Specifes study as 'classif
                                     icaton variable'         */
model or=/ s cl;                  /* Intercept only model, show
                                     solution and CI          */
repeated /group =study;           /* Each trial has its own
                                     within trial variance    */
parms /parmsdata=sillar           /* The parmsdata option reads
                                     in the variable EST indi
                                     cating the variances from
                                     the data set sillar.sd2 */
eqcons=1 to 36;                   /* The within study variances
                                     are known and fixed      */
run;
```

C. SAS Code for the Random Effects Model

The SAS procedure proc mixed requires the following manipulations of the data

```
data covvars;     /* data set containing the variances  */
set sillar;
keep est;
run;
data start;          /* include the starting value for the  */
input est;           /* heterogeneity variance              */
cards;
0.0
run;
data start;          /* Combine both data sets             */
set start covvars;
run;
```

Obtain the model with proc mixed

```
proc mixed method=ml cl data=sillar; /* CL asks for confidence intervals             */
                                     /* of covariance parameters                     */
class study;                         /* Study is classification variable             */
model or= / s cl;                    /* Intercept only model, Fixed solution and CI  */
random int /subject=study ;          /* Study is specified as random effect          */
repeated /group =study;              /* Each study has its own variance              */
parms /parmsdata= start              /* start contains starting value a. trial vars. */
eqcons=2 to 37;                      /* entries 2 to 37 are the fixed study vars.    */
run;
```

References

Allam MF, Del Castillo AS, Navajas RF (2003) Parkinson's disease, smoking and family history: meta-analysis. Eur J Neurol 10:59–62

Baker SG, Kramer BS (2002) The transitive fallacy for randomized trials: if a bests b and b bests c in separate trials, is a better than c? BMC Med Res Methodol 2(1):13

Begg CB, Mazumdar M (1994) Operating characteristics of a rank correlation test for publication bias. Biometrics 50:1088–1101

Bennett DA (2003) Review of analytical methods for prospective cohort studies using time to event data: single studies and implications for meta-analysis. Stat Methods Med Res 12:297–319

Berlin JA (1995) Invited commentary: benefits of heterogeneity in meta-analysis of data from epidemiologic studies. Am J Epidemiol 142:383–387

Berlin JA, Begg CB, Louis TN (1989) An assessment of publication bias using a sample of published clinical trials. J Am Stat Assoc 84:381–392

Berlin JA, The University of Pennsylvania Meta-Analysis Blinding Study Group (1997) Does blinding of readers affect the results of meta-analyses? Lancet 350:185–186

Blair A, Burg J, Floran J, Gibb H, Greenland S, Morris R, Raabe G, Savitz D, Teta J, Wartenberg D, Wong O, Zimmerman R (1995) Guidelines for application of meta-analysis in environmental epidemiology. Regul Toxicol Pharmacol 22:189–197

Blettner M, Sauerbrei W, Schlehofer B, Scheuchenpflug T, Friedenreich C (1999) Traditional reviews, meta-analyses and pooled analyses in epidemiology. Int J Epidemiol 28:1–9

Boffetta P (2002) Involuntary smoking and lung cancer. Scand J Work Environ Health 28(Suppl 2):30–40

Boffetta P, Saracci R, Andersen A, Bertazzi PA, Chang-Claude J, Cherrie J, Ferro G, Frentzel-Beyme R, Hansen J, Plato N, Teppo L, Westerholm P, Winter PD, Zochetti C (1997) Cancer mortality among man-made vitreous fiber production workers. Epidemiology 8:259–268

Böhning D, Dietz E, Schlattmann P (1998) Recent developments in computer assisted mixture analysis. Biometrics 54:283–303

Böhning D, Malzahn U, Dietz E, Schlattmann P, Viwatwongkasem C, Biggeri A (2002) Some general points in estimating heterogeneity variance with the DerSimonian-Laird estimator. Biostatistics 3:445–457

Boyd NF, Martin LJ, Noffel M, Lockwood GA, Trichler DL (1993) A meta-analysis of studies of dietary fat and breast cancer. Br J Cancer 68:627–636

Bucher HC, Guyatt GH, Griffith LE, Walter SD (1997) The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 50(6):683–691

Caldwell DM, Welton NJ, Ades AE (2009) Mixed treatment comparison analysis provides internally coherent treatment effect estimates based on overviews of reviews and can reveal inconsistency. J Clin Epidemiol. doi:10.1016/j.jclinepi.2009.08.025

CGHFBC Collaborative Group On Hormonal Factors Breast Cancer (1996) Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies. Lancet 347:1713–1727

Cook DJ, Guyatt GH, Ryan G, Clifton J, Buckingham L, Willan A, McLlroy W, Oxman AD (1993) Should unpublished data be included in meta- analyses? current conflictions and controversies. J Am Med Assoc 21:2749–2753

Cooper NJ, Sutton AJ, Morris D, Ades AE, Welton NJ (2009) Adressing between-study heterogeneity and inconsistency in mixed treatment comparisons: application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. Stat Med 28(14):1861–1881

Copas JB, Shi JQ (2001) A sensitivity analysis for publication bias in systematic review. Med Res 10:251–265

DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. Control Clin Trials 7:177–188

Dickersin K (2002) Systematic reviews in epidemiology: why are we so far behind? Int J Epidemiol 31:6–12

Dickersin K, Scherer R, Lefebvre C (1994) Identifying relevant studies for systematic reviews. BMJ 309:1286–1291

Egger M, Davey Smith G, Schneider M, Minder C (1997) Bias in meta-analysis detected by a single, graphical test. BMJ 315:629–634

Egger M, Schneider M, Davey Smith G (1998) Spurious precision? meta-analysis of observational studies. BMJ 316:140–144

Egger M, Davey Smith G, Altman DG (2001a) Systematic reviews in health care. Meta-analysis in context, 2nd edn. BMJ Publishing Group, London

Egger M, Davey Smith G, Schneider M (2001b) Systematic reviews of observational studies. BMJ Publishing Group, London, pp 211–227

Egger M, Juni P, Bartlett C, Holenstein F, Sterne J (2003) How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? An empirical study. Health Technol Assess 7:1–76

Elliott WJ, Meyer PM (2007a) Incident diabetes in clinical trials of antihypertensive drugs: a network meta-analysis. Lancet 369(9557):201–207

Elliott WJ, Meyer PM (2007b) Incident diabetes in clinical trials of antihypertensive drugs. Lancet 369:1514–1515

Feinstein AR (1995) Meta-analysis: statistical alchemy for the 21st century. J Clin Epidemiol 48:71–79

Friedenreich CM (1993) Methods for pooled analyses of epidemiologic studies. Epidemiology 4:295–302

Friedenreich C (1994) Influence of methodologic factors in a pooled analysis of 13 case-control studies of colorectal cancer and dietary fiber. Epidemiology 5:66–67

Glass GV (1977) Integrating findings: the meta-analysis of research. Rev Res 5:3–8

Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, Damico R, Bradburn M (2005) Indirect comparisons of competing interventions. Health Technol Assess 9(26):1–134

Greenland S (1987) Quantitative methods in the review of epidemiologic literature. Epidemiol Rev 9:1–302

Greenland S (1994) Invited commentary: a critical look at some popular meta-analytic methods. Am J Epidemiol 140:290–296

Hamajima N, Hirose K, Tajima K (2002) Alcohol, tobacco and breast cancer–collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease. Br J Cancer 87:1234–1245

Higgins JPT, Green S (2009) Cochrane handbook for systematic reviews of interventions, Version 5.0.2. The Cochrane Collaboration, London

Higgins JP, Whitehead A (1996) Borrowing strength from external trials in a meta-analysis. Stat Med 15(24):2733–2749

Hill AB (1965) The environment and disease: association or causation? Proc R Soc Med 58:295–300

Hill HA, Kleinbaum DG (2000) Bias in observational studies. Wiley, Chichester, pp 94–100

Hung RJ, Boffetta P, Brockmoller J (2003) CYP1A1 and GSTM1 genetic polymorphisms and lung cancer risk in caucasian non-smokers: a pooled analysis. Carcinogenesis 24:875–882

Jones DR (1992) Meta-analysis of observational epidemiologic studies in a consistent form. J R Soc Med 85:165–168

Longnecker MP, Berlin JA, Orza MJ, Chalmers TC (1988) A meta-analysis of alcohol consumption in relation to risk of breast cancer. JAMA 260:652–656

Lu G, Ades AE (2004) Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med 23(20):3105–3124

Lu G, Ades AE (2006) Assessing evidence inconsistency in mixed treatment comparisons. J Am Stat Assoc 101(474):447–459

Lubin JH, Boice JD Jr, Edling C, Hornung RW, Howe G, Kunz E, Kusiak RA, Morrison HI, Radford EP, Samet JM, Timarche M, Woodward A, Yao SX (1995) Radon-exposed underground miners and inverse dose-rate (protraction enhancement) effects. Health Phys 69:494–500

Lumley T (2002) Network meta-analysis for indirect treatment comparisons. Stat Med 21(16):2313–2324

Macaskill PS, Walter SD, Irwig L (2001) A comparison of methods to detect publication bias in meta-analysis. Stat Med 20:641–654

Meinert R, Michaelis J (1996) Meta-analyses of studies on the association between electromagnetic fields and childhood cancer. Radiat Environ Biophys 35:11–18

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009) Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. The Cochrane Collaboration. PLoS Med 6(7):e1000097. doi:10.1371/journal.pmed.1000097

Nixon RM, Bansback N, Brennan A (2007) Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis. Stat Med 26(6):1237–1254

Normand SL (1999) Meta-analysis: formulating, evaluating, combining, and reporting. Stat Med 18:321–359

Olkin I (1994) Invited commentary: Re: "a critical look at some popular meta-analytic methods". Am J Epidemiol 140:297–299

Paul SR, Donner A (1989) A comparison of tests of homogeneity of odds ratios in k 2x2 tables. Stat Med 8:1455–1468

Peters J, Sutton A, Jones D, Abrams K, Rushton L (2006) Comparison of two methods to detect publication bias inmeta-analysis. The Cochrane Collaboration. JAMA 295(6):676–680

Pettiti DB (1994) Meta-analysis, decision analysis and cost-effectiveness analysis. Oxford University Press, New York

Morris RD (1994) Meta-analysis in cancer epidemiology. Environ Health Perspect 102:61–66

Riley RD, Simmonds MC, Look MP (2007) Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. J Clin Epidemiol 60(5):431–439. doi:10.1016/j.jclinepi.2006.09.009

Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, Jackson RD, Beresford SA, Howard BV, Johnson KC, Kotchen JM, Ockene J, Writing Group for the Women's Health Initiative Investigators (2002) Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. JAMA 288:321–333

Salanti G, Higgins JP, Ades AE, Ioannidis JP (2008) Evaluation of networks of randomized trials. Stat Methods Med Res 17(3):279–301

Sauerbrei W, Blettner M, Royston P (2001) Letter to White, IR (1999): "The level of alcohol consumption at which all-cause mortality is least". J Clin Epidemiol 54:537–538

Schlattmann P (2009) Medical applications of finite mixture models. Statistics for biology and health. Springer, Berlin, 246p. EUR 64.15

Schlattmann P, Böhning D (1993) Computer packages c.a.man (computer assisted mixture analysis) and dismap. Stat Med 12:1965

Schlehofer B, Blettner M, Preston-Martin S, Niehoff D, Wahrendorf J, Arslan A, Ahlbom A, Choi WN, Giles GG, Howe GR, Little J, Menegoz F, Ryan P (1999) Role of medical history in brain tumour development. Results from the international adult brain tumour study. Int J Cancer 82:155–160

Schwarzer G, Antes G, Schumacher M (2002) Inflation of type I error rate in two statistical tests for the detection of publication bias in meta-analyses with binary outcomes. Stat Med 21:2465–2477

Shapiro S (1994a) Is there is or is there ain't no baby? Dr. Shapiro replies to Drs. Petitti and Greenland. Am J Epidemiol 140:788–791

Shapiro S (1994b) Meta-analysis/shmeta-analysis. Am J Epidemiol 140:771–778

Sidik K, Jonkman JN (2005) Simple heterogeneity variance estimation for meta-analysis. Appl Stat 54:367–384

Sillero-Arenas M, Delgado-Rodriguez M, Rodiguesw-Canteras R, Bueno-Cavanillas A, Galvez-Vargas R (1992) Menopausal hormone replacement therapy and breast cancer: a meta-analysis. Obstet Gynecol 79:286–294

Smith ML, Glass GV (1977) Meta-analysis of psychotherapy outcome studies. Am Psychol 32(9):752–760

Smith-Warner SA, Ritz J, Hunter DJ, Albanes D (2002) Dietary fat and risk of lung cancer in a pooled analysis of prospective studies. Cancer Epidemiol Biomark Prev 11:987–992

Song F, Altman DG, Glenny AM, Deeks JJ (2003) Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. BMJ 326(7387):472

Song F, Harvey I, Lilford R (2008) Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions. J Clin Epidemiol 61(5):455–463

Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DG (2009) Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. BMJ 338:b1147. doi:10.1136/bmj.b1147

Steinberg KK, Smith SJ, Lee N, Stroup DF, Olkin I, Williamson GD (1997) A comparison of meta-analysis to pooled analysis: an application to ovarian cancer. Am J Epidemiol 145:1917–1925

Stock WA (1995) Systematic coding for research synthesis. The Russell Sage Foundation, New York, pp 1–2

Straif K, Chambless L, Weiland SK, Wienke A, Bungers M, Taeger D, Keil U (1999) Occupational risk factors for mortality from stomach and lung cancer among rubber workers: an analysis using internal controls and refined exposure assessment. Int J Epidemiol 28:1037–1043

Straif K, Keil U, Taeger D, Holthenrich D, Sun Y, Bungers M, Weiland SK (2000) Exposure to nitrosamines, carbon black, asbestos, and talc and mortality from stomach, lung, and laryngeal cancer in a cohort of rubber workers. Am J Epidemiol 152:297–306

Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, Moher D, Becker BJ, Sipe TA, Thacker SB (2000) Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis of observational studies in epidemiology group. J Am Med Assoc 283:2008–2012

Stukel TA, Demidenko E, Dykes J, Karagas MR (2001) Two-stage methods for the analysis of pooled data. Stat Med 20:2115–2130

Sutton AJ, Higgins JP (2008) Recent developments in meta-analysis. Stat Med 27(5):625–650

Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F (2000) Methods for meta-analysis in medical research. Wiley, Chichester/New York

Sutton AJ, Kendrick D, Coupland CA (2008) Meta-analysis of individual- and aggregate-level data. Stat Med 27(5):651–669

Thompson SG (1994) Why sources of heterogeneity in meta-analysis should be investigated. BMJ 309:1351–1355

Thompson SG, Sharp SJ (1999) Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med 18:2693–2708

Thompson SG, Smith TC, Sharp SJ (1997) Investigating underlying risk as a source of heterogeneity in meta-analysis. Stat Med 16:2741–2758

Tweedie RL, Mengersen KL (1995) Meta-analytic approaches to dose-response relationships, with application in studies of lung cancer and exposure to environmental tobacco smoke. Stat Med 14:545–569

Ursin G, Longenecker MP, Haile RW, Greenland S (1995) A meta-analysis of body mass index and risk of premenopausal breast cancer. Epidemiology 6:137–141

van Howelingen HC, Arends LC, Stijnen T (2002) Advanced methods in meta-analyis: multivariate approach and meta-regression. Stat Med 59:589–624

Weed DL (1997) Methodologic guidelines for review papers. JNCI 89:6–7

Weed DL (2000) Interpreting epidemiological evidence: how meta-analysis and causal inference methods are related. Int J Epidemiol 29:387–390

Whitehead A (2002) Meta-analysis of controlled clinical trials. Wiley, Chichester

Zeeger MP, Jellema A, Ostrer H (2003) Empiric risk of prostate carcinoma for relatives of patients with prostate carcinoma: a meta-analysis. Cancer 97:1894–1903

Geographical Epidemiology 37

John F. Bithell

Contents

J.F. Bithell
St Peter's College, University of Oxford, Oxford, UK

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_22,

37.1 Introduction

37.1.1 The Nature of Geographical Epidemiology

Although, at first sight, geographical epidemiology may appear to differ substantially from other areas of epidemiology, it has many features in common. In particular, a major objective of epidemiology – to infer etiological relationships from observed associations – applies also in geographical studies. The distinctive characteristic is of course that geographical location is an important explanatory variable, either because it reflects an environmentally determined element of risk or because people with similar risk attributes live together, so that risk varies from place to place. The two-dimensional nature of geographical location means that the standard statistical techniques for handling sets of essentially univariate variables need to be augmented by more sophisticated methods.

There are practical limitations to the scientific value of geographical studies. The data quality tends to be low – not least because population censuses are relatively infrequent – and any real effects may be attenuated by factors such as mobility, often to the point where they are not detectable. Consideration of these difficulties may lead to the conclusion that a lot of geographical epidemiology is, in scientific terms, of very limited value. Historically, however, there have been some spectacular successes: to the famous observation of Snow (1855) on the source of cholera infection may be added a number of more recent and equally dramatic observations, for example, the identification of the cause of an outbreak of asthma in Spain (Antó and Sunyer 1992) and the implication of erionite fibres in the etiology of mesothelioma from the very high localized rates in the Cappadocian region of Turkey (Baris et al. 1992).

37.1.2 Scope of the Chapter

This chapter attempts to sketch the statistical principles of the subject, with an indication of the kinds of analyses to which these principles lead quite naturally. There is a large literature on the methodology of geographical epidemiology, much of it employing a Bayesian standpoint and exploring hierarchical models analyzed by Markov chain Monte Carlo methods. It would be impossible to give a comprehensive review of the latter field, and we adopt the less ambitious objective of outlining the fundamentals of the subject, in the hope that this will in any case provide insight into more sophisticated analyses. Nevertheless we have attempted to provide some examples of the techniques discussed and, where possible, to make recommendations for practitioners, though this latter goal is difficult in view of the large number of different analyses that have been proposed but whose properties are relatively unknown.

Our presentation will in fact be almost exclusively frequentist. To some extent, the choice between Bayesian and frequentist methods in statistics is a matter of philosophical standpoint. Frequentist arguments are undeniably limited in their scope and power and are frequently subject to misinterpretation. The limitations may, however, be argued to be intrinsic to the problem of inductive inference under uncertainty and such inference does not seem to this author to be more consistently clear-cut when derived from a Bayesian analysis. The modeling approach is admittedly more attractive than the mere detection of statistical significance, but it is not without its difficulties. For one thing, the amount of data in geographical studies may often not permit the estimation of numerous parameters, and to the extent that a model makes specific assumptions about underlying phenomena, there is a risk that it may inject spurious information into the analysis, leading to the overinterpretation of the data. The limitations of the hypothesis testing approach have not prevented its widespread use in practice, and an important part of the epidemiologist's role is to ensure that the tests that are carried out are chosen with due regard to maximizing their power against sensible alternatives. This at least is the standpoint from which we approach this topic here; in any case, the statistical framework underpins the more sophisticated analyses and forms a natural prerequisite for their understanding. See chapter ▸Statistical Inference of this handbook for a discussion of the fundamental distinctions between Bayesian and frequentist inference, and chapter ▸Bayesian Methods in Epidemiology of this handbook for an account of Bayesian modeling.

37.1.3 Chapter Contents

We start by considering (Sect. 37.2) the models that underlie statistical methods in geographical epidemiology in order to give insight into the justification for the methods that are discussed. A key feature is the duality that exists between the two approaches to epidemiological investigations generally. To be specific, we can elect to study either the occurrence of disease conditionally on case locations or vice versa, i.e., to regard case location as a random variable to be compared in fixed groups of affected and unaffected individuals. This duality precisely mirrors the distinction between the cohort and case-control approaches to epidemiological surveys. The case-control approach in geographical work has only recently been recognized and is particularly relevant for the analysis of data at the individual, as opposed to the areal, level. This important approach, though not yet fully exploited, has led recently to a number of new and interesting methodological developments.

In Sect. 37.3, we develop the way in which risk may be modeled in relation to geographically referenced data, distinguishing between the analysis of areal data and data at the individual level, for which it is assumed that individual locations are known. As with any statistical modeling exercise, the objective is to explain as much of the variation as possible, up to the point where heterogeneity can be attributed to chance. There are numerous ways of approaching this subject, even within the compass of frequentist analyses, and some of the issues as to the best analysis are unresolved.

Section 37.4 is concerned with mapping. From one point of view, mapping is an end in itself, and there are numerous methods available for producing maps. However, there is much scope for misinterpretation of data represented in this way, and we would argue that a map should be seen as the end product of some kind of modeling process, though possibly a very primitive one: no disease map can be constructed without assumptions about the underlying distribution of the disease it purports to represent.

Section 37.5 addresses the question of heterogeneity in the distribution of risk. To some extent, this involves issues bound up with the problems of modeling. But the simple question of whether there is any non-uniformity of risk is a valid one that can be at least partially answered without reference to underlying models or alternatives.

In Sect. 37.6, we address the problem of clustering. This may be seen as a violation of the twin assumptions of uniformity and independence discussed in Sect. 37.2. However, we may well be more interested in detecting small clusters of cases that are related to one another, and to this extent it may be appropriate to use different methods from those in Sect. 37.5.

Finally, Sect. 37.7 considers the rather more specific problem of detecting an increase in risk near a putative point source of risk, and it is argued that analyses of this kind are essentially one dimensional, and perhaps for this reason, it is somewhat easier to determine good methods for doing so. This is in fact a problem of considerable interest, and many investigations of "clustering" are really of this kind. The issue is illustrated by the incidence of childhood leukaemia around nuclear installations in the UK using data introduced in Sect. 37.3.2.

The concluding section summarizes the chapter and makes suggestions for further reading.

37.2 Statistical Models

In this section, we describe a statistical framework for the methods to be discussed. We start by explaining the elements that underlie the analysis of classical surveys and then show how the same starting point may be applied to geographical data.

37.2.1 A Statistical Framework for Epidemiological Observations

To describe a modeling framework for epidemiology, we start by supposing that the disease $\mathcal{D}$ in which we are interested is an essentially dichotomous entity, i.e., it is the binary outcome – affected/not affected – of some biological process applied to a finite set of individuals. Such a starting point will serve irrespective of the temporal nature of the events we are studying, be they deaths or incident cases of a disease $\mathcal{D}$ in a given time period or the prevalence of $\mathcal{D}$ at a given epoch. We will be primarily interested in the association between $\mathcal{D}$ and various covariates $\mathcal{C}$. Some of these may represent risk factors suspected of playing a causal role: we will describe these as exposure variables and denote them by $\mathcal{E}$. Others may be of interest in their own right or because they are potential confounding variables for $\mathcal{E}$. We will treat $\mathcal{E}$ as a subset of $\mathcal{C}$ when this is convenient.

To take a specific geographical example, we cite the famous study of cardiovascular disease $\mathcal{D}$ by Cook and Pocock (1983). The covariates $\mathcal{C}$ included water hardness $\mathcal{E}$, whose etiological relationship to cardiovascular disease was of primary interest, and also various indicators of socioeconomic status, which played the role of a confounding factor: the gradients of mortality, water hardness, and socioeconomic status are highly correlated with latitude in the UK. The data were analyzed for males and females together, but they could equally well have been stratified by sex, which would be a covariate of interest in its own right, since one might be interested in the mortality of males and females separately.

Next, we assume that occurrences of $\mathcal{D}$ are independent. This does not preclude the possibility that individuals have probabilities p of $\mathcal{D}$ that are related through their proximity, for example. Rather the condition stipulates that, *conditional on the values of* $\mathcal{C}$ *and* $\mathcal{E}$, the occurrence of $\mathcal{D}$ in one individual is independent of that in another, i.e., that the probability that individual A suffers from $\mathcal{D}$ is unaffected by the *fact* (as opposed to the probability) that some other individual B also suffers from it. In practice, this is a reasonable mechanistic assumption for nearly all chronic disease epidemiology. It clearly breaks down for infectious diseases, for which more sophisticated models would be appropriate. In fact, little theoretical foundation for modeling the epidemiology of infectious diseases at the individual level exists. This is partly because the theory is intractable, partly because it is not necessary in setting up the null hypothesis of no contagion for the purposes of testing. It is only for formulating alternative hypotheses in this situation that statistical models for a contagious mechanism are necessary. Important though this is, we will not consider the problem in this chapter.

Under this independence assumption, the individual outcomes of $\mathcal{D}$ are described by the very simple Bernoulli distribution. If all the probabilities p_i for the individuals in a group of n are the same, the number of occurrences out of the n will clearly follow the binomial distribution, while if all the p_i are different and supposed to depend on $\mathcal{C}$, we can model them through a (binary) logistic regression (Cox and Snell 1984).

Such analyses are becoming more common, but they require detailed information on individuals and are not without their technical difficulties. Much of epidemiology is in practice still conducted by the more traditional approach of grouping data according to disease status and to grouped values of $\mathcal{C}$. In this approach, the assumption is that the probabilities p_i within a particular group are indeed all the same, though in practice we know that this is unlikely to be true. However, this assumption is far less troublesome than appears at first sight. For one thing, as long as the p_i are small, the difference between a binomial distribution and that of a sum of slightly different Bernoulli variables will be negligible.

A typical analysis of epidemiological data proceeds by forming a cross-tabulation into a contingency table, whose rows, columns, and layers are labeled by components of $\mathcal{D}$, $\mathcal{E}$, and $\mathcal{C}$. The standard way of analyzing such a table is through a log-linear model, which implicitly assumes that the counts in the table are values of Poisson distributed variables, conditioned by the requirements that certain subtotals in the table are deemed to be fixed. For details on log-linear regressions please refer to chapter ▸Regression Methods for Epidemiological Analysis of this handbook.

The logistic regression and log-linear modeling approaches thus described have been constructed on the assumption that $\mathcal{D}$ is a random response and the covariates $\mathcal{C}$ are fixed, but we can also obtain useful analyses by conditioning on the numbers of "cases" affected by $\mathcal{D}$ and unaffected or disease-free "controls" in a suitable control group and regarding one or more of the covariates as a random response. This leads to the so-called "case-control" study (formerly termed a *retrospective* study), in distinction to a "cohort" (or *prospective*) study. Thus, for example, it might be appropriate to use a normal linear regression to model the exposure $\mathcal{E}$ of individuals to some risk factor – considered to be a continuous variable – as a function of the other variables, one of which would be an indicator for $\mathcal{D}$, the membership of the case or control group. We would then regard $\mathcal{E}$ as the factor of primary interest and the other covariates would be fitted in order to control for their possible confounding effects.

37.2.2 Statistical Models for Geographical Data

Most of the ideas outlined above carry over quite naturally to data in which geographical location plays a role. We will preserve the assumptions that $\mathcal{D}$ is a binary variable and that disease occurrences are independent conditionally on $\mathcal{C}$. We need to extend our conceptual notation to include geographical location, which we will denote by $\mathcal{G}$. There is a distinction between situations where we think of it as representing a pair of coordinates and those where it is an essentially two-dimensional location in the space representing a geographical region studied.

If $\mathcal{G}$ is thought of as representing coordinates, such as Easting and Northing, it may be meaningful to treat them like other quantitative variables, perhaps to detect a trend with latitude, for example. Alternatively, it might be meaningful to consider polar coordinates from a specified point $\mathcal{S}$ considered as a fixed origin, analyzing distance and direction from it. Typically, $\mathcal{S}$ would be a point of some etiological significance, such as a putative source of pollution. We return to this topic in Sect. 37.7 below.

However, this approach implicitly reduces our analyses to consideration of essentially one-dimensional variables, and it is useful to distinguish this from the *intrinsically spatial* case in which we regard two-dimensional space as a single entity. In this situation, a principal objective will be to depict the way in which risk varies over a region $\mathcal{R}$, usually by means of a map. It is unlikely that any kind of analytically determined trend surface, such as a polynomial, will be useful, though non-parametrically estimated surfaces might be. We return to the problems of mapping in Sect. 37.4 below.

The distinctions we made in Sect. 37.2.1 above apply for geographical data. For example, the majority of geographical analyses are effectively analyses of grouped data, in which observations have been grouped into k subregions $A_1, A_2, \ldots, A_k$ of

$\mathcal{R}$ (which we shall refer to as "areas"). Within each area, we would hope to know the population to serve as a denominator and the number of occurrences Y_i of the disease $\mathcal{D}$ would then follow a binomial or approximately a Poisson distribution, by the arguments outlined above. The areas may be regarded as analogous to the bins of a histogram, though they will nearly always be based on administrative areas with highly irregular boundaries, so that they do not share the attractive regularity properties of the more familiar histograms formed from quantitative variables. The identities of the areas themselves typically enter the analysis through the coordinates of their population centroids, and these may then be analyzed by incorporating them into the model as described above, though the analysis might well take account of spatial autocorrelation.

If instead of binning or grouping the observed cases into areas we record the exact locations of the occurrences of $\mathcal{D}$, we need a rather different modeling approach. The case-independence assumption implies that the cases are located according to a non-homogenous Poisson process (Diggle 2000), which is the standard probability model for events happening at random in a continuum, though not necessarily with a uniform pattern of risk. This model supposes that the probability of an event in a small area δA at the point (x, y) is $\lambda(x, y)\delta A$, where $\lambda(x, y)$ is the "intensity function" of the process giving the rate per unit area at (x, y); it also incorporates the crucial assumption that the occurrence of such a point is independent of occurrences outside δA.

It is well known, however, that when points occur according to a Poisson process in such a way that the total number is fixed at a value n, say, the pattern of points obtained is typically exactly the same as if we had sampled from a probability distribution with density function proportional to $\lambda(x, y)$. This enables us to describe the behavior in geographical space of a fixed sample of cases, with a view to estimating the risk at each point (x, y) or to compare the resulting risk function with that for a sample of controls. Thus we have moved to the "dual" or case-control approach, for we are effectively regarding the locations as realizations of a continuous bivariate random variable defined for our samples of cases and controls. Methods of analyzing data within this framework are discussed in Sect. 37.3.4 below.

37.3 Modeling Disease Risk in Relation to Geographically Referenced Factors

37.3.1 Areal Data

One of the commonest and most straightforward analyses of geographical data consists of modeling the counts Y_i of cases in areas A_i using a Poisson regression or, equivalently, a generalized linear model (GLM) with Poisson error and log link function; see McCullagh and Nelder (1989) and chapter ▶Regression Methods for Epidemiological Analysis of this handbook. We start by assuming that we can calculate “null expectations” e_i for the Y_i. In the simplest form, these could be obtained by multiplying some global reference estimate of risk p by the population sizes in the A_i. In practice, we will almost certainly wish to *standardize* for the age distribution and other known demographic factors such as socioeconomic status. Part of our objective is of course to modify the assumption that the risk is the same in every area, so we will incorporate a *relative risk* (*RR*) θ_i, to give the model for the counts as

$$Y_i \sim \text{Poisson}[\theta_i e_i] \; .$$

We then model the θ_i in the usual manner for a GLM through

$$\log \theta_i = \sum_{j=1}^{p} x_{ij} \beta_j \; ,$$

where the β_j are coefficients in the log-linear model and x_{ij} is the value of the jth covariate for the ith areal unit A_i.

Typical covariates in such an analysis might include intrinsically geographical features, such as altitude, geological composition, or levels of background radiation, or essentially demographic features, such as the age or socioeconomic composition of the population of each area. It should be emphasized that the units in such analyses are not the individuals with a disease $\mathcal{D}$ but the areas within which they reside, and the covariates are also necessarily attributes of these areas. The object of such an analysis, however, will generally be to make inferences concerning individuals, and to ignore the distinction is sometimes described as perpetrating the “ecological fallacy.” Covariate values for the area as a whole are implicitly imputed to each individual member of the population, and this has the potential for introducing a number of different kinds of bias known variously as “ecological” or “aggregation bias.”

A genuinely “ecological” or geographical imputation would arise if a geographical feature (such as latitude) were averaged spatially without regard to population density (Diggle and Elliott 1995), and any such averaging should as far as possible be density-weighted, perhaps by using the relevant measurement at the centroid of the population. Demographic variables, such as age or socioeconomic deprivation will usually be averaged over the population anyway, and in this case grouping into areas has much the same effect as grouping by other factors; the problem may then be seen in the wider context of aggregation bias. Such a bias can result from concealed within-group confounding, and it is difficult to take account of this without individual-level data. An intrinsic bias may also arise from the non-linearity of the model used, though this is likely to be small when the disease risk is itself small and relatively uniform, since the logistic (or any similar) transformation used in the model will be nearly linear.

Many papers have addressed the issue of ecological bias. An early contribution by Greenland and Morgenstern (1989) was influential but may have painted too pessimistic a view of ecological studies, which can be very valuable for providing pointers and which are often based on more objective data than case-control studies. Wakefield (2008) provides a useful overview and a review of the literature. Further discussion of the issues will also be found in chapter ▸Descriptive Studies of this handbook.

37.3.2 An Example of the Log-Linear Model for Areal Data

An example of this use of the log-linear model is provided by the application to childhood leukaemia data described by Bithell et al. (1995). The dataset analyzed was from the UK National Registry of Childhood Tumours (NRCT) maintained by the Childhood Cancer Research Group in Oxford and related to 5,359 children diagnosed with leukaemia or non-Hodgkin lymphoma under the age of 15 years between 1966 and 1987. Each of the cases was located in one of 9,831 electoral wards, which are administrative areas with an average population of around 5,000.

The explanatory variables fitted were "Standard Region," a classification of Britain into ten regions, and the Townsend Index, an areal index of social deprivation which is a function of unemployment, housing ownership, and other socioeconomic indicators. As shown in Table 37.1, there was a significant reduction in the deviance associated with each of these factors: the p-values shown in the first two lines are based on the chi-square approximation to the deviance reductions. It is interesting, incidentally, to note that the direction of the association is negative for the Townsend Index, i.e., the disease is slightly commoner in less deprived families. This is a feature of childhood leukaemia that differentiates it from most other diseases.

The goodness of fit of the model can in principle be tested by the residual deviance, but because the expected numbers of cases per ward in this analysis were small (less than half on average), the chi-square approximation is unreliable. However, the theoretical mean and variance of the deviance for Poisson observations with a specified set of expectations can be calculated straightforwardly. We can therefore obtain an approximate test of the residual deviance as follows:

1. Compute the values for the expectations predicted by the model for each ward.
2. Compute the mean μ and variance σ^2 of the deviance statistic D as defined by

$$D = 2\sum_i [Y_i \log(Y_i/e_i) - (Y_i - e_i)] \tag{37.1}$$

   as if the contributions to D were independent.
3. Refer the statistic $(D - \mu)/\sigma$ to the standard normal distribution.

Table 37.1 Analysis of deviance of childhood leukaemia data

Variation due to	d.f.	Deviance	p-value
Standard region	9	23.1	0.0060
Townsend index	1	23.6	10^{-6}
Residual	9,820	8610.6	0.025

The assumption of independence should be approximately true in view of the large number of degrees of freedom. Bithell et al. check the two-sided p-value by simulation of data from the fitted model and found a very good degree of approximation to the calculated value of 0.025. These results may be interpreted as meaning that the model fits much better than if the explanatory factors had not been taken into account (for which the equivalent p-value was 0.00042); though there is some evidence of residual heterogeneity, it must be remembered that this is a large data set and the level of significance observed is not indicative of a large degree of variation. We return to the issue of testing residual variation in Sect. 37.5.

37.3.3 Calculating the Expectations

The model described above involves the expectations e_i, which appear as an "offset" term in the model, i.e., $\log(e_i)$ is added to the linear function of the covariates defining $\log \theta_i$. These may be calculated from externally calculated rates, for example, from national statistics. If such rates are not easily available, the data can be internally standardized by supplying the sizes of the populations at risk; any factor representing the overall risk will appear in the intercept term of the model. The expectations predicted by the model can then be used as expectations for subsequent analyses, and this is a useful by-product of the modeling process. The method can be seen as an elegant and more consistent alternative to classical standardization, permitting the flexible inclusion of covariates according to their importance, as indicated by the modeling process.

Indeed, the analysis described by Bithell et al. is part of a larger one designed to produce expected numbers of childhood leukaemias for the areal analysis of incidence near nuclear installations; this is briefly described in Sect. 37.7.2.

37.3.4 Continuous Data

Following the discussion in Sect. 37.2.2 above, we suppose that we have a sample of exact locations of cases of disease $\mathcal{D}$ and that we denote their density function over $\mathcal{R}$ by $\psi(x, y)$. We need an analogue of the denominators in an areal analysis to serve as a measure of how many individuals there are at risk at each point (x, y) of $\mathcal{R}$. This is provided in principle by knowledge of the population density, which we will consider to be continuous and which we will denote by $\pi(x, y)$. Then our problem becomes one of comparing the density function for the incident cases with that of the population. For a rare disease, the population density (which strictly speaking includes diseased as well as healthy individuals) will be very similar to that for all non-diseased individuals, which can in turn be estimated by a suitable *sample* of controls. The natural way to make this comparison is through the ratio, and it is easily seen that this ratio

$$\theta(x, y) = \psi(x, y)/\pi(x, y)$$

defines a *relative risk function* (RRF) that gives the risk of being affected by $\mathcal{D}$ at each point (x, y) of $\mathcal{R}$ relative to the mean for the whole of $\mathcal{R}$ (see Bithell 1990).

A natural estimate $\widehat{\theta}(x, y)$ of $\theta(x, y)$ is provided by the ratio of estimates of $\psi(x, y)$ and $\pi(x, y)$. These may be obtained using one of the modern methods available for estimating a density function (see the books by Silverman (1986) and Scott (1992), e.g.). The process is not without difficulties but it can be used to provide meaningful estimates of the RRF over $\mathcal{R}$, in effect providing a map of it. We return to the problem of mapping in Sect. 37.4 below.

A more ambitious objective than merely mapping the RRF is to model it as a function of covariates $\boldsymbol{x}$, say. These may be geographically defined at every point of $\mathcal{R}$, or they may be attributes of the cases and controls in the samples. An elegant modeling approach is due to Diggle and Rowlingson (1994) and proceeds by analogy with classical case-control studies. We condition on the coordinates of the n cases and m controls and consider the probability that, under random allocation of the cases and controls to the $m + n$ locations, an individual sampled at a given location (x, y) is a case rather than a control. This probability can then be modeled logistically as a function of $\boldsymbol{x}$. If there appears to be unexplained variation in the RRF, it can in principle be modeled by adding a non-parametric function of (x, y) to the linear predictor. The numerical problems of the latter approach appear not to be trivial.

The inclusion of attributes of the individuals in the analysis is particularly attractive, since it provides the possibility of controlling for them within the geographical analysis. In practice, it is not always straightforward to obtain suitable controls for analyses of this kind, partly because the current emphasis on data protection makes it difficult to access individual records and partly because of the number of combinations of categories with respect to which we may wish to match. Nevertheless, this methodology, though still in its infancy, would seem to have considerable potential.

37.3.5 Spatial Structure in the Residual Variation

The object of fitting a model of the kind discussed is to obtain a satisfactory explanation of the data, i.e., a residual deviance that is not statistically significant. This is not always very easy, since the risk of disease may depend on factors that we have been unable to measure. Large data sets – for example, of national mortality rates – may also demonstrate a statistically significant deviance resulting from unobserved factors that are scientifically unimportant simply because of the large numbers of cases involved.

Unfortunately, conclusions about the importance of individual explanatory variables in a model are strictly valid only if the model fitted is correct. In practice, we will believe a model to be correct if it appears to fit reasonably well, i.e., if the residual deviance is not statistically significant. This raises the question of how to proceed if there is a degree of residual variation that we cannot explain.

In geographical studies, it is quite likely that such variation will be due to unobserved variables that are spatially autocorrelated, and in this case we can include terms in the model designed to reflect this autocorrelation. Typically, this is done for data in areal form using a conditional autoregression (CAR) model (Wakefield et al. 2000) while, for continuous data, Kelsall and Diggle (1998) use a generalized additive model (GAM) which effectively gives an extra term in the model estimating residual variation non-parametrically. These ideas are important but are somewhat beyond the scope of this chapter; see Pfeiffer et al. (2008) for an introductory account of spatial models and Diggle (2000) for a good overview of the field. We only remark that the issue may not always be as significant as some authors maintain. The deviances of the terms that are fitted in a model will still be a reliable indication of their importance unless they are confounded with the unobserved variables that are inflating the deviance; in this case, fitting a spatial model merely tells us that this confounding has a spatial structure – it does not help us to identify the variable or determine its scientific importance.

37.4 Mapping Disease Risk

The mapping of disease risk is a central endeavor of geographical epidemiology: a map is as convenient for portraying such location-specific information as it is for indicating the geography of the land to which it relates. It is therefore no surprise to discover that mapping has a long history predating any systematic development of the statistical principles that underlie it.

As with other areas of geographical epidemiology, many methods have been proposed. Broadly speaking, these can be divided into two classes, model-based and non-parametric. Methods in each of these classes can be applied to data in either areal or continuous form. It is important to appreciate, however, that, whatever method is applied, there is inevitably a degree of smoothing involved that is to some extent arbitrary and under the control of the investigator.

For example, the simplest form of map is the so-called choropleth map, which uses a gray or color scale to depict the risk of $\mathcal{D}$ in each of a number of areas, usually administratively defined so that denominators are easily available. Here, the degree of smoothing is determined by the size of the areas A_i, since the process represents the risk as being the same throughout each given area. An example of a choropleth map is given in chapter ▸Descriptive Studies of this handbook.

Similarly, data in continuous point form can be mapped using the methods described in Sect. 37.3.4 by plotting the RRF $\widehat{\theta}(x, y)$. Here, the smoothing is determined by the degree of smoothing used in the estimation of the densities: it is a commonplace of this methodology that some smoothing parameter always has to be used, though there are data-driven methods for estimating the most appropriate value. See Bithell (1990) for an early example of this method applied to small numbers of cases and controls, and Davies and Hazelton (2010) for a more recent development of the methodology.

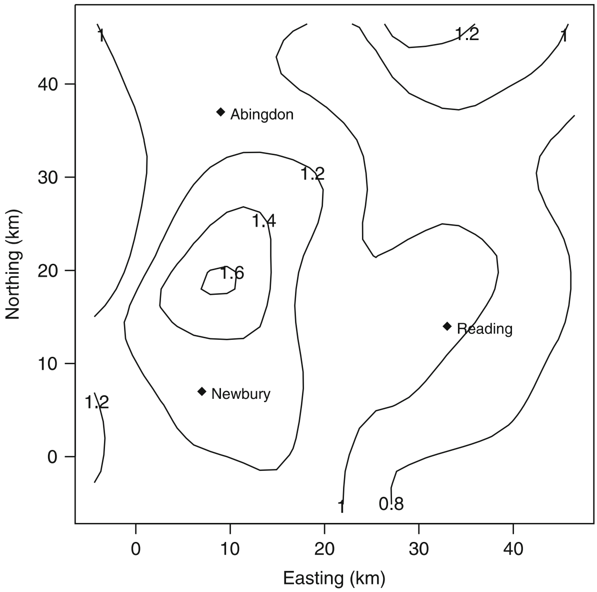


Fig. 37.1 Relative risk function for childhood cancer in a region of Oxfordshire, estimated from areal data. The three town centers are shown only approximately. The ASH smoothing parameter used was 8 (See Bithell 1999 for details)

It may be noted that the RRF method can easily be adapted to areal data by suitably modifying the customary density estimation methods (Bithell 1999). Figure 37.1 depicts the incidence of childhood cancer in a 50-km square region of Oxfordshire using data from the UK National Registry of Childhood Tumors maintained by the Childhood Cancer Research Group in Oxford. They consist of 279 cases of childhood cancer (other than leukaemia and non-Hodgkin lymphoma) registered under the age of 15 years between 1966 and 1987. Each case was located in one of 150 electoral wards for which expected numbers of cases were calculated using similar methods to those for the leukaemia data described in Sect. 37.3.2. The point observations for the cases were used to construct a density estimate $\widehat{\psi}(x, y)$ using the *average shifted histogram* (ASH) method due to Scott (1992). For the controls, the density estimate $\widehat{\pi}(x, y)$ was constructed by treating the centroids of the wards as point locations weighted by the expectations and using a version of ASH modified accordingly.

The basis of the ASH method is to count the numbers of cases in the cells of a square grid; these are then smoothed by slightly shifting the grid a number of times and averaging the resulting counts; this process effectively smoothes the surface by spreading out the contributions of the points through neighboring grid squares.

The RRF was then obtained by dividing the density estimates for the cases and controls to give $\widehat{\theta}(x, y) = \widehat{\psi}(x, y)/\widehat{\pi}(x, y)$. This is depicted in Fig. 37.1 as a contour plot with a scale in km and an origin located in South West Oxfordshire.

The methods sketched above may be regarded as empirical or non-parametric, in that there is nothing underlying them that is more sophisticated than the division of one number by another (specifically a count by a denominator or one density estimate by another). It is generally difficult to see how to determine the appropriate degree of smoothing by any objective process, as distinct from using intuitively plausible and aesthetically pleasing values.

The necessity for a degree of smoothing can easily be seen by considering a choropleth map, for which we have areas A_i with small numbers of cases, either because we have chosen small areas or because $\mathcal{D}$ has low incidence. In this case, the estimates of the risk in each A_i will be subject to large sampling errors; our belief about the true risk in A_i will be determined in part by the observed rate, but it will also rely on information from the region as a whole, to the extent that we believe there will be some comparability between the areas.

This idea has led to the development of model-based approaches using Bayesian arguments to integrate area-specific information with information from the whole region, using a statistical model for the underlying variation of the true risk. In a classical treatment of this problem, Clayton and Kaldor (1987) suppose that the true risk θ_i in A_i is distributed over the areas as a whole according to a gamma distribution with mean μ and variance σ^2. It can then be shown that the posterior distribution of θ_i has mean

$$\tilde{\theta}_i = \frac{y_i + \mu^2/\sigma^2}{e_i + \mu/\sigma^2},$$

where y_i is the observed value of the count Y_i in A_i. This formula can be seen to be a form of average of the maximum likelihood estimate $\widehat{\theta_i} = y_i/e_i$ of each θ_i and the overall mean μ, which can be estimated by $\sum y_i / \sum e_i$. The value of σ^2 can also be estimated from the data as a whole, though this requires an iterative method.

This method and variants of it provide *empirical Bayes* estimates, in that the prior distribution of the θ_i can be estimated from the data. The method is essentially non-spatial, in the sense that the true θ_i is supposed to vary independently. In practice, it is likely that rates in neighboring areas will be consistently more similar to one another than those in more separated areas. If this were not so, it would be essentially fruitless to attempt to produce a smoothly varying map. The Bayesian methodology has been extended to permit the prior distribution of the θs to depend on the values in neighboring areas. These more complicated models involve a greater number of arbitrary assumptions, however. They are gaining ground in popularity and appear to be used quite successfully. The reader is referred for more details and references to Clayton and Bernardinelli (1992) and to chapter ▸Bayesian Methods in Epidemiology of this handbook.

Attractive though these ideas are, the maps they produce need careful interpretation, since they have imposed a degree of spatial autocorrelation, and this process is capable of making adjacent areas look more similar than they really are. In a sense, this is true of all mapping methods and is a feature as intrinsic as the implicit smoothing itself.

In a challenging paper, Gelman and Price (1999) discuss the issue and illustrate the phenomenon of induced spatial pattern by means of simple modeling paradigms.

They point out that the probability that a particular area rate $\widehat{\theta_i}$ exceeds a given value increases with decreasing population size, n_i, say. The effect of this is that high observed rates of disease tend to be observed predominantly in low population areas; since these tend to be spatially aggregated – i.e., low population areas are more likely to occur next to other such areas – observed rates also appear to be spatially related even when in fact no such relationship exists for the underlying risk.

They further demonstrate that plotting the posterior means from a Bayesian analysis produces observed rates that are likely to exceed a particular value with probabilities that are *decreasing* functions of n_i, so that such plots overcorrect in some sense. Although scores exist – at least for continuous observations – that are not subject to these artifacts, they have no direct interpretation as estimates of the θ_i.

One is driven to the conclusion that disease maps are potentially misleading when used as anything except what Gelman and Price call "look-up tables," i.e., as a convenient way of depicting the rate in a given area *without reference to neighboring areas*. It is the temptation to use the map to generalize about the spatial pattern of rates that can be misleading, and it is probably better to formulate such questions within the context of a statistical model rather than to attempt to portray spatial relationship graphically. However, one suspects that this timely caution is unlikely to diminish the enthusiasm for constructing and overinterpreting disease maps.

37.5 The Detection of Generalized Heterogeneity

37.5.1 The Assessment of Heterogeneity in Areal Data

Heterogeneity is the key to epidemiology, in the sense that a uniform risk in observed data gives no possibility for associating differences with factors that may have etiological significance. We have already touched on the issue of modeling in Sect. 37.3.1, and our objective there is to find a model that appears to fit well in the sense that the residual deviance is not statistically significant – i.e., it is consistent with chance deviations from the predictions of the model.

As long as we have Poisson data with reasonably large means, we can assess the residual deviance as if it had a chi-square distribution with a number of degrees of freedom (d.f.) determined by the model – specifically the number of units minus the number of parameters fitted. It is important to remember, however, that this is based on asymptotic theory which, roughly speaking, supposes that the total number of cases is large compared with the number of units – areal or otherwise – in the analysis. A rule of thumb suggests that the expectations of the counts in a Poisson regression should mostly be in excess of 5. When the average expectation falls below this, we should expect the distribution of the deviance in a correct model to depart progressively from a chi-square distribution, which of course means that a corresponding statistical test of goodness of fit of the model based on the chi-square distribution would not be valid.

In this situation, we can obtain an approximate assessment of the value of the deviance – and hence the goodness of fit of the model – by simulation. Typically, we would generate, say, s new samples of data from Poisson distributions with means obtained from the model $\mathcal{M}_{\text{fitted}}$ fitted to the actual data. For each simulated sample, we would refit the same model and compute the residual deviance. The s values of the deviance thus obtained provide an estimate of the distribution of the deviance. This in turn provides a means of calibrating the deviance observed for our actual data. A formal test of goodness of fit would only be approximate since we are simulating from $\mathcal{M}_{\text{fitted}}$ rather than the true model with the true (unknown) parameter values. This situation is typical of "bootstrapping," and the theory of this subject could in principle lead to better approximations. For an account of bootstrapping see, for example, Efron and Tibshirani (1993).

37.5.2 Detecting Heterogeneity in Poisson Data

A special case arises when we have expectations, provided, for example, by some prior analysis or by simple calculation from population data and we merely wish to detect whether the Poisson distribution fits well with the assumed e_i, without reference to any model fitting. This is sometimes seen as a problem of detecting "clustering," though there are qualifications to this interpretation that we discuss below: for the moment, we prefer to regard this as the problem of assessing heterogeneity, i.e., variations in risk between areas without reference to a possible geographical origin for the phenomenon.

Relating this to the deviance of a Poisson model suggests that the deviance of the observations, defined in Eq. 37.1, Sect. 37.3.2, would be a sensible test statistic. The fact that this test is a likelihood ratio test means that it is *asymptotically* fully efficient – i.e., its power approaches that of the best possible test against a Poisson alternative in which the relative risks are different from unity.

Popular alternative contenders include Pearson's chi-square statistic

$$X^2 = \sum (Y_i - e_i)^2 / e_i \ ,$$

and the Potthoff–Whittinghill statistic (Potthoff and Whittinghill 1966)

$$PW = \sum Y_i (Y_i - 1) / e_i \ ,$$

which is regarded by some authors as a test of clustering. The former is, at least in simple cases, asymptotically equivalent to the deviance but is easier to compute and to study analytically. The asymptotic requirement, however, implies that the

Table 37.2 Expected significance levels (ESL) % and their standard errors for Pearson's X^2, the deviance, and the Potthoff–Whittinghill tests: k wards each with expectation e under H_0 and an alternative expectation dispersion with variance σ^2

e	σ^2	k		Pearson	Deviance	Potthoff
5.0	0.05	200	ESL	6.6	3.5	44.5
			s.e.	0.22	0.18	0.50
1.0	0.2	500	ESL	3.5	2.7	6.3
			s.e.	0.18	0.16	0.24
0.2	1.0	1,000	ESL	14.2	23.8	3.3
			s.e.	0.35	0.43	0.20

expectations should be large and the theoretical properties give rather little guidance on which test is best for small expectations.

Table 37.2 shows the results of a simulation study, designed to provide such guidance, in which the expected significance level (ESL) of each test has been estimated in each of three conditions. (The ESL is a convenient alternative criterion to power (Dempster and Schatzoff 1965): a smaller ESL corresponds to a more powerful test.) In each case, the ESLs were estimated from 10,000 simulations performed under varying conditions. These were chosen to produce values in a critical range corresponding to situations where the test would be quite likely to lead to different conclusions at conventional significance levels. In each case, a specific number k of wards were supposed to have the same expectations e under the null hypothesis, while under the alternative hypothesis, these expectations were multiplied by a set of RR factors θ_i sampled from a gamma distribution with mean one and variance σ^2.

In interpreting this table, we suppose that the key parameter is the size of the expectation e. Because the test statistic will be roughly proportional to the number of wards k, this latter parameter represents the amount of information and was chosen to bring the ESLs into an interesting range; it would not be expected to change the relative ordering of the three tests. The variance σ^2 represents the distance between the null and alternative hypotheses, and the values were chosen to be typical of the sort of discrepancy that one could reasonably expect to detect in practical situations. It could conceivably affect the relative properties of the different tests but is less likely to do so than e.

It will be seen that, with an expectation of $e = 5$, the deviance is indeed the best test, while the Potthoff–Whittinghill test trails behind Pearson's chi-square test. The difference between X^2 and D becomes marginal around $e = 1$ while, for smaller expectations, the ordering is reversed and the Potthoff–Whittinghill test appears to be superior. These results suggest that it would be wise to carry out simulations in particular marginal cases to determine the best test to use. It should also be emphasized that one should evaluate the significance of the chosen statistic using simulation when the e_i are small, since the Pearson and deviance statistics are then likely to have distributions markedly different from the chi-square.

37.5.3 Spatial and Non-Spatial Analyses

A test of heterogeneity in areal data of the kind described above provides only a non-spatial test of the heterogeneity of our observations. Whether this is appropriate depends on whether or not the areal units are defined by essentially geographical criteria. If, for example, they are defined by simply dividing our region $\mathcal{R}$ into urban and rural areas, then a factor associated with the degree of urbanization could be expected to induce heterogeneity into the areas irrespective of their spatial positions.

More frequently, however, areas are merely convenient administrative subdivisions of $\mathcal{R}$. In this case, we might expect a factor that raises the incidence in one area to do so in adjoining areas also. Then, a test that takes no account of the spatial relationship of the areas will be less powerful than one that does.

To take a simple hypothetical example, suppose that $\mathcal{R}$ consists of two subregions: $\mathcal{R}_1$ with n areas each having expectation $e_i = 9$ and $\mathcal{R}_2$ with n areas each having expectation $e_i = 11$. A dispersion test based on Pearson's chi-square statistic would use the variance of the observations to test the null hypothesis H_0 that all the expectations are the same:

$$X^2_{2n-1} = \sum_{i=1}^{2n} (Y_i - e)^2 / e \, ,$$

where $e = \sum_1^{2n} Y_i / n$ is the (estimated) expected count based on all $2n$ observations. To a good approximation, this statistic would have a chi-square distribution with $2n - 1$ degrees of freedom under H_0. If, however, we knew which areas belonged to $\mathcal{R}_1$ and which to $\mathcal{R}_2$, we would base the test on the equivalent statistic for testing the difference between the totals for the two subregions:

$$X^2_1 = \frac{\left(\sum_1^n Y_i - ne\right)^2 + \left(\sum_{n+1}^{2n} Y_i - ne\right)^2}{ne},$$

and it is fairly obvious that this would be a much more powerful test of H_0. This idealized situation is analogous to isolating sources of variation in an analysis of variance.

In practice, of course, we will almost certainly not be in a position to divide $\mathcal{R}$ into high- and low-risk areas a priori, but this example does suggest that the detection of non-uniformity of risk should take account of the spatial structure of the data. A classical account of tests of spatial autocorrelation is given by Cliff and Ord (1981), who establish some theoretical properties of their sampling distributions, particularly in the case of normally distributed observations. In one of the few comparative studies published, Walter (1993) examines the power empirically for three of the most popular tests against a variety of geographically plausible alternatives. The three considered were as follows:

- The I statistic of Moran (1948), which is analogous to a correlation coefficient and is defined by
$$I = \frac{n \sum_{ij} w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{ij} w_{ij} \sum_i (x_i - \bar{x})^2}$$
- The c statistic of Geary (1954), which is similar to I
- A non-parametric test statistic which uses only the ranks of the observations

The first two statistics used as observations $x_i = y_i/e_i$, the standardized incidence ratios for the different areas, and spatial weights w_{ij} chosen to be one if A_i and A_j are adjacent and zero otherwise. Walter's Table II shows that, in each of the situations he considered, Moran's I had the highest power of the three, and it would seem that this should be the method of choice, at least for detecting generalized spatial relationship as opposed to isolated peaks in the risk. The question of whether higher power could be achieved by using more sophisticated weighting than a simple adjacency matrix, or by weighting the pairs of observations according to the amount of information they contain (in terms of sample size, for example), has not been much considered. Walter concludes that "the precise type of spatial pattern involved may have a major impact on the spatial power of the analysis" and that "more experience is needed to better understand the potential of these methods, and their limitations." Nevertheless this study was a useful contribution, and the use of Moran's I to detect spatial autocorrelation is probably a good choice.

37.5.4 Heterogeneity Tests Based on the Risk Surface

If we have continuous data – i.e., observations at the individual level – we can base a test of uniformity on the RRF $\widehat{\theta}(x, y)$ as estimated by the methods described in Sect. 37.3.4. We may regard a test statistic as being defined by a functional of $\widehat{\theta}(x, y)$, and there are various possibilities.

A natural choice is the weighted variance of $\widehat{\theta}(x, y)$:

$$T_{\text{var}} = \iint_{\mathcal{R}} \pi(x, y)\{\widehat{\theta}(x, y) - 1\}^2 \mathrm{d}x\mathrm{d}y \ .$$

In the absence of any reliable theory, it is necessary to resort to Monte Carlo methods to test the statistic. For case-control data, we use a permutation method that is straightforward though laborious:

1. Construct a map of the risk function $\theta(x, y)$ by a suitable method, using a degree of smoothing which is determined as a function of the data.
2. Evaluate the chosen test statistic for the observed data t_{obs}.
3. Choose a new sample of "cases" by choosing at random n points from the set of $m + n$ cases and controls combined.
4. Compute the value of the statistic t_1, say, for the simulated data, using the same procedure as in step 1.

5. Repeat steps 3 and 4 $s - 1$ more time so that there are s simulated values altogether.
6. Reject at level $\alpha = m/(s+1)$ the null hypothesis of uniformity of cases if t_{obs} is greater than all but $m - 1$ of the simulated observations.
7. Alternatively estimate the p-value of the test as the number of $\{t_i \geq t_{\text{obs}}\}/s$.

This general Monte Carlo procedure is applicable in very general circumstances, and it is especially useful in the analysis of spatial data, where construction of suitable models is difficult. We must remember, however, that a hypothesis test is, by itself, of very little inferential value without some idea of how probable the observed results would be under a plausible alternative.

The method can easily be adapted to a test based on a risk surface constructed from areal data as described in Sect. 37.4. The simulation would take the form of sampling areal counts from Poisson distributions with expectations e_i and computing the variance over a square grid as before. In either the continuous or areal data case, the degree of smoothing used in the density estimation process determines the scale of aggregation for which the test is most sensitive and is analogous to the choice of weights w_{ij} in Moran's statistics.

The use of tests of this sort is still in its infancy, but the underlying philosophy is attractive and increasing computing power is making them more practicable even for large data sets.

37.6 Clustering

Closely related to the idea of heterogeneity is the concept of clustering, with which much of geographical epidemiology is preoccupied. There is a large literature on the subject, not all of which is very clear on the issue of what we actually mean by the words "cluster" and "clustering." We may conveniently define a cluster as a localized aggregation of disease cases greater than can easily be explained by chance. Clustering may be regarded as the tendency to form clusters or, more generally, as any departure from the assumptions of uniform risk and independence of case occurrences as discussed in Sect. 37.2.1. We will continue to use the word heterogeneity to refer to a departure from uniformity and reserve the word clustering as far as possible to refer to mechanisms in which case occurrences are not independent. This kind of clustering may be supposed to act locally, whereas heterogeneity is more likely to be observed throughout $\mathcal{R}$ and is sometimes referred to as "generalized clustering." For further discussion of the issues the reader is referred to a useful paper by Diggle (2000).

We can give here only the briefest of accounts. We will distinguish between methods based on increased levels of risk and methods based on the proximity of neighbors. First, however, we make two general points about clustering.

In the first place, it is a well-accepted fact of spatial statistics that it is not possible to distinguish on the basis of a single realization of observed data from a spatial process whether any non-uniformity of the distribution of points (relative to an expected population distribution) is due to a variation of underlying risk, with cases occurring independently (i.e., points generated by a non-homogeneous Poisson process or its equivalent), or to a mechanism in which existing cases induce others nearby, such as would happen in a contagious process. Secondly, we remark that, from an abstract point of view, clustering may take place in any continuum and, in the geographical context, we may observe clustering in space, time, or the "product-space" of time and geographical space. This mathematical commonality means that tests can be adapted from one problem to another, with very fruitful consequences.

37.6.1 Methods Based on the RRF

Clustering is likely to be observed as an increase in risk in some locality and it follows that we can use the estimated risk surface $\widehat{\theta}(x, y)$ to provide an appropriate test. What functional of $\widehat{\theta}(x, y)$ we use will depend on the alternative we have in mind or, equivalently, the pattern we would most like to detect. If, for example, we are content to demonstrate a single cluster or aggregation of cases we could choose as our test statistic, the maximum of the $\widehat{\theta}(x, y)$ over the whole region $\mathcal{R}$:

$$T_{\max} = \max_{x,y \in \mathcal{R}} \{\widehat{\theta}(x, y)\} .$$

This does not, of course, preclude the possibility that we would detect multiple clusters, but it is likely that our test would be most powerful in the situation where there are in fact very few. We could of course extend the statistic to consider, for example, the mean of the r largest peaks in $\widehat{\theta}(x, y)$, but it is unlikely that we would have good a priori grounds for fixing r. Tests based on peaks of incidence must also be expected to be quite sensitive to the scale of the clustering phenomenon and to the degree of smoothing we employ in constructing $\widehat{\theta}(x, y)$.

A statistic likely to have similar properties to $T_{\max}$ is based on a scanning window, typically a square that moves over $\mathcal{R}$. At each point of a fine grid the observed number of cases is compared with its expectation; the test statistic is defined as the maximum discrepancy using a suitable criterion such as the incidence ratio. Here, the size of the window plays the role of a smoothing parameter; the main difference from $T_{\max}$ is that a peak incidence is weighted according to its radial extent; it seems likely that it behaves in a similar manner to $T_{\max}$ for suitably chosen smoothing parameters. Anderson and Titterington (1995) describe a version of this method that varies the window size to keep constant the expected number of cases under the null hypothesis. Much subsequent work describing similar tests has been published; see Tango (2010) for a recent summary. Some of these have considered windows of different shapes, notably elliptical, but the usefulness of these is limited, not least because we are unlikely to know a priori what shape a cluster might have.

In fact, the scanning window is a two-dimensional version of an approach originally used for detecting clustering in time; even this one-dimensional version is notoriously intractable analytically and simulations or other numerical methods would seem to be unavoidable.

37.6.2 Knox's Test

The use of what we may call pairing methods is historically older than the methods based on the risk surface discussed above; they have the attraction of being very simple to describe and understand.

The earliest such test is due to Knox (1964), who counted the number, Z, of pairs of children with leukaemia diagnosed within 60 days and 1 km of each other in Northumberland and Durham, two counties in the North East of England (see Table 37.3, taken from Knox (1964)). The study used local registration and hospital records, as well as death certificates to ascertain 185 children with an onset of leukaemia under 15 years of age between the years 1951 and 1960 inclusive. However, certain cases were excluded, and Table 37.3 refers just to children under the age of six, a restriction that needs to be borne in mind when interpreting the results; in fact, older children showed no effect.

The rationale for this test is explicitly related to the non-independence of the cases, namely, that a contagious mechanism passing a disease from one individual to another would be likely to lead to cases that are closer to one another in space and time than would be expected by chance. This in turn leads to the idea of considering pairs of cases.

Knox refers this statistic to its expectation calculated on the assumption that the spatial locations and times of occurrence of the disease are independent. This is given by

$$\mathbf{E}[Z] = \frac{N_T N_S}{\binom{n}{2}} \,,$$

where N_T, N_S are the numbers of pairs of cases close in time and close in space, respectively, and the denominator is the total number of pairs out of the n cases.

In effect, this becomes a test of the independence of these two variables, and it uses their *marginal* distributions to determine the null distribution of Z. Knox conjectures that Z should follow a Poisson distribution approximately; this is shown to be true in certain circumstances in work reported by David and Barton (1966), who give a formula for the variance of Z. It is wise to calculate this or to use a Monte Carlo test in which the times of occurrence of the cases are randomly permuted relative to the space coordinates, and the statistic Z is recomputed a large number of times. For Knox's data, the value of $\mathbf{E}[Z]$ is 0.83, for which $Z = 5$ has a p-value of 0.0017 when tested as a Poisson observation. David

Table 37.3 Pairs of cases of childhood leukaemia classified according to their closeness in space and time (see text)

		Distance apart (km)		
		0–1	Over 1	Total
Time apart	0–59	5	147	152
(days)	60–3,651	20	4,388	4,408

and Barton report an early simulation experiment for Knox's data carried out by M.C. Pike; the latter finds $Z \geq 5$ in 4 out of 2,000 simulations. This leads to an estimated significance level of 0.002 which is very close to that based on the Poisson approximation.

The choice of the critical distance and time separation is of course crucial. It determines the scale of clustering likely to be detected, and it should ideally be fixed in advance for the formal validity of the testing procedure. In particular, it is certainly not formally valid to test at a large number of different critical distances and times and then select the most significant result without allowance for this selection. If we really have no idea of the time and distance scales that would be appropriate, we need to use a data-driven method of identifying the most promising values (see Sect. 37.6.6).

37.6.3 Other Space-Time Clustering Methods

An alternative test based on the proximity in space and time of pairs of cases is proposed by Jacquez (1996). This is based on the number out of the l nearest neighbors in space of a given case that are also among the l nearest neighbors in time. Like the Knox test, it can be adapted to provide a test of space-only clustering. Jacquez claimed superior power to that of the Knox test, though in practice, this is likely to depend on the alternative being considered. Here, the parameter l serves as a kind of scale parameter since it determines how far we look for association between cases.

Knox's very elegant idea permits us to dispense with the need to estimate the marginal distributions, though only under the assumption that space and time are in fact independently distributed in the population. This assumption applies of course to Jacquez' test also. It will clearly be violated by population drift, i.e., a change of population distribution with time. Kulldorf and Hjalmars (1999) examine the size of this effect and conclude that it can be "a considerable problem." They recommend that space-time clustering should be tested using the joint space-time distribution of the population size, but this is of course rather hard to obtain with good accuracy and resolution. It seems likely that the use of the interaction tests will remain popular.

37.6.4 Space-Only Clustering

Knox's idea of counting pairs has been very fruitful and has been adapted to a number of related situations, including the use of a sample of controls to provide a reference distribution when testing for space-only clustering (Pike and Smith 1974). The essential idea here is to regard the controls as being similar to the cases, except that they are considered to have occurred at different "pseudo-times," while the cases are considered to have occurred simultaneously. The statistic computed is then the number of pairs of cases that are close in space, and it is not hard to see that this is formally equivalent to Z, with identical distributional statistical properties. Knox's test is not the only test that can be adapted to detecting space-only clustering using controls. Other possibilities are explored by Rogerson (2006) in a paper giving some analytical results but no power study.

37.6.5 Population Distance

One alternative to this adaptation of Knox's test for case-control data is a kind of dual approach due to Cuzick and Edwards (1990). This is based on the count of the number of individuals among the l nearest neighbors of each case that are also cases (as opposed to controls). The quantity l in the Cuzick–Edwards test serves as a determinant of the scale of clustering to be detected in this method. It is given in terms of the number of individuals likely to be within a region of contagion, rather than by a distance.

This may be seen as more relevant for some, though not for all, mechanisms of disease spread. Indeed, for any given pair, we can think of closeness in terms of distance or in terms of the number of other members of the population residing between the two members of the pair. The choice between these two metrics is crucial, though which is the more appropriate will presumably depend on the supposed etiology of the disease.

The idea of a population distance lies behind another method of testing, due to Besag and Newell (1991), who consider each case in turn and aggregate the areas around it that are necessary to include the rth nearest case. The expectation for the aggregate of these areas is then compared with r in the usual way. This can be regarded as a kind of inverse sampling, and again the number of cases considered, r, is a parameter that determines the scale of clustering to which the procedure is sensitive.

37.6.6 Choosing Scale Parameters

Every clustering phenomenon has an implied scale of the clustering effect and it is clearly desirable to have some idea of this before attempting to detect it. When we have no idea, the temptation to perform multiple testing arises and it is important to make allowance for this. A method for testing a range of distances and times in the Knox test is proposed by Abe (1973); effectively, this examines a multi-way table for association between space and time, making due allowance for the non-independence of the pairs. This statistic is sensitive to association over the whole range of distances and times rather than attempting to identify the most interesting scale. To identify the scale of maximal clustering effect, we can use a general data-driven procedure that can be constructed along the following lines:

1. Test the data at each of a number of critical space and time distance pairs.
2. Form a single test statistic, either using some aggregate over different values of the scale parameters or using some measure of the maximum degree of clustering; call this statistic t_{obs}.
3. Simulate further data sets under the null hypothesis: for Poisson data, this will probably involve sampling Poisson-distributed counts, while for case-control data, it may involve pooling all the cases and controls and randomly selecting a subset to serve as simulated "cases."
4. Rank the simulated values of the statistic $t_1, t_2, \ldots, t_s$ and compare the ranked values with t_{obs}.

This Monte Carlo procedure is of general applicability and provides a way of getting round the problem of unknown scale. It does of course sacrifice power by comparison with a test that correctly focuses on the true degree of clustering, so that the more carefully alternative hypotheses can be framed a priori the better.

Faced with this wide variety of tests, it is difficult for the researcher to know which to use. Each new test published typically is claimed to be more powerful than previously existing tests, but there is a wide variety of alternatives to uniformity of risk that could be considered, and it is certain that no one test is uniformly most powerful against all alternatives. In principle, it is open to the researcher to examine competing tests to see which would be best for the data and the alternative hypothesis in question, but this can be an arduous exercise. This is an area where we badly need more insight into which tests are preferable.

37.7 Predefined Sources of Risk

One of the epidemiological questions most often asked in a geographical context is whether there appears to be an aggregation of cases around a putative source of risk $\mathcal{S}$ such as an industrial plant. For example, there has been much interest in the UK, as in other countries, in the possibility of an elevated risk of childhood leukaemia around nuclear power stations. This results in part from the finding of an unusually large aggregation of cases near the nuclear reprocessing plant at Sellafield, which is situated on the coast of Cumbria in the North West of England. In fact, ordinary nuclear generating stations have little in common with the reprocessing plant and the experimental reactor at Sellafield, nor is there evidence of significant releases of radioactivity into the environment from generating stations. Nevertheless, public anxiety persists about the safety of the plants, partly perhaps because of the difficulty in comprehending the nature of nuclear power and partly because of sensational reporting in the news media. In fact, there is little evidence of a general increase in risk (Bithell et al. 1994), but it is highly desirable that the best statistical procedures are used to test the data that come under scrutiny. The public may not have a very sophisticated understanding of statistics, but it is obvious even to the uninitiated that some of the procedures used in the past have not been well-chosen from the point of view of maximizing the chance of detecting a real effect.

Aggregations around $\mathcal{S}$ are sometimes referred to as "clusters," but it is not generally supposed that the cases involved are related, only that the risk to individuals in the vicinity of $\mathcal{S}$ is elevated. Analyses could therefore proceed using the methods described in Sect. 37.3.1, with the obvious qualification that geographical variables clearly represent spatial relationship to $\mathcal{S}$. In practice, this nearly always means using distance from $\mathcal{S}$ or some function of it, so that the analysis is implicitly one-dimensional. Moreover, analyses are often required in situations where the number of cases is very small, and in this situation, the fitting of GLM's tends to be unstable and to lead to parameter estimates with large standard errors and unknown distributional properties.

37.7.1 Tests for Concentration of Risk

In this situation, it is probably better to rely on a formal significance test and the issue then becomes that of selecting the most powerful test against a suitable hypothesis or range of hypotheses. The resulting analyses are likely not to be very powerful in any case, but choosing the most powerful test at least increases the chance that a significant result can be attributed to a genuine departure from the null hypothesis of uniform risk.

The method of early investigators of simply comparing the risk in the area around $\mathcal{S}$ with a reference or "control" rate outside the area defines a test procedure that is in fact powerful only against an alternative hypothesis prescribing a uniform excess risk within the area, which drops to zero on the boundary. This is clearly implausible and critically dependent on the size of the area chosen; one inevitably concludes that a better test would be one designed for some systematic relationship between the risk and the distance from $\mathcal{S}$. We may reasonably suppose that this relationship is monotonic, but the rate of decay and the shape of the RRF (expressed now as a function of distance) will determine the power of the test.

An ingenious class of tests designed to be powerful against general monotonic alternatives was proposed by Stone (1988). His "MLR test" compares the ratio of the maximum of the likelihood under the null hypothesis of uniform risk against the likelihood of the observations maximized subject to the restriction that the risk is a non-increasing function of distance from $\mathcal{S}$, i.e.,

$$H_1 : \theta_1 \geq \theta_2 \geq \ldots \geq \theta_k \quad (\geq 1) \,, \tag{37.2}$$

where θ_i is the relative risk in the ith area in order of increasing distance from $\mathcal{S}$. Stone's test has become very popular in the UK epidemiological literature, though it is known that it is never the most powerful test against a specific hypothesis, this being provided by a *linear risk score* (LRS) test of the form

$$T = \sum_j \ln\left(\theta\left(d_j\right)\right) \,,$$

where d_j is the distance of the jth case from $\mathcal{S}$ and $\theta(d)$ is the risk at a distance d from $\mathcal{S}$ as specified by the alternative hypothesis (Bithell 1995).

Unfortunately, knowing the most powerful test against a specific alternative hypothesis does not greatly help if we do not know what that alternative is. However, it provides a benchmark against which we can judge other tests, and in particular, it enables us to determine the sensitivity of the power to variations in the alternative. It turns out that statistics of the form

$$T = \sum_j 1/\phi\left(d_j\right) ,$$

for monotonic functions $\phi(\cdot)$, define a class of *canonical* tests which can come close to optimal power in many circumstances. In particular,

$$\phi\left(d_j\right) = d_j \text{ and } \phi\left(d_j\right) = \sqrt{\text{rank}\left(d_j\right)}$$

behave well in areas with a reasonably uniform population distribution. However, the latter affects quite strongly which of the canonical tests actually is most powerful and it is wise to check the performances of the competing tests in each different study using simulation. In addition to their simplicity, the canonical tests have the great advantage that they are not dependent on any parameters in the RRF; the test based on the reciprocal of distance, for example, is most powerful against all alternatives for which the RRF is of the form $a \exp(b/d_j)$, for any parameters a (which governs the overall degree of risk) and b (which governs the rate of decay). The fact that risk is unbounded at zero is a small price to pay for this advantage, which means that there is no need to perform multiple tests with different values of a and b.

Because the LRS test statistics are sums, they should in principle have an approximately normal distribution, and it is easy to compute their moments. In small samples, this asymptotic normal approximation will not necessarily apply, and it is advisable to use simulation also to carry out the tests, i.e., to carry out Monte Carlo tests. In doing so, it is easy to see that the way the samples are drawn can be either to fix the total number of cases and use the multinomial distribution or to use unconstrained Poisson distributions to determine the counts in the areas A_i. Which of these two sampling schemes is used is very important and will typically affect the results quite substantially. The first method defines a conditional test which might be appropriate if the expectations e_i for the rates in the different areas are unreliable in absolute terms (though possibly still all right relatively); it is important to note though that, if the expectations are correct, a conditional test could reject the null hypothesis because of a deficit of cases near the boundary of the region rather than an excess near $\mathcal{S}$. The second, unconditional, test is appropriate if the rates are reliable and in this case the test statistic combines the evidence from the overall relative risk in the area with that from the spatial distribution. In this case, the appropriate form of Stone's test should also include the last (bracketed) inequality in Eq. 37.2 above.

Many other tests have been proposed for testing the concentration of risk around a point source; these are sometimes referred to as "focused tests." Some of these are in the class of LRS tests, though this is not always recognized. Some have been designed to use polar coordinates (Lawson 1993) and so to test for a directional effect; unfortunately no equivalent of the canonical test appears to be available for this problem, and so the maximal direction of the effect is a nuisance parameter that has to be estimated from the data unless there is a clear a priori reason for choosing one specific direction. Such tests have been applied to rather few datasets in practice. Tango (2010) gives a recent review of the field and reports extensive power calculations; these confirm that the canonical LRS tests do reasonably well for non-directional alternatives corresponding to a smooth monotonic RRF.

37.7.2 Example: Childhood Leukemia Around UK Nuclear Installations

The tests described above were developed partly in conjunction with analyses of the distribution of childhood leukaemia around nuclear installations. An analysis of all major sites in England and Wales is described by Bithell et al. (1994) using the data on leukaemia and non-Hodgkin lymphoma described in Sect. 37.3.2. The sites were examined separately using the LRS test with the reciprocal of distance rank as the primary test, though Stone's MLR test was also used for comparison. As remarked above, the results were largely negative.

However, public interest in the possibility of a raised risk persists, and two subsequent updated unconditional analyses have been published (COMARE 2005, 2011). In the first of these, the analyses were carried out in the light of a large simulation study that identified which of a number of tests would be most powerful at each of the sites. Experience of these analyses suggests that the power does indeed depend on the population distribution, but it has been found that, for the majority of test sites studied, the most powerful test against the alternatives considered was the LRS test based on $1/\sqrt{\text{distance rank}}$.

Table 37.4 shows the average power averaged over 75 alternative hypotheses and the significance levels achieved by each of five tests for one of the datasets from the 1994 analysis. It will be noticed that the smallest p-value was the Poisson maximum (often known as "Pmax"); this is in effect the maximum value of the cumulative relative risk as we move out from $\mathcal{S}$. The most powerful test, on the other hand, gives a non-significant result. This analysis is a timely warning against judging a test by the significance level achieved in a real dataset. More details and discussion of this analysis are given in Bithell (2003).

In the later study of nuclear power plants in Britain (COMARE 2011), the analysis was restricted to children under the age of 5 years and distances closer to the installations. This reduced the numbers to the point where it was necessary

Table 37.4 Average power of five tests and significance levels achieved for the 80 wards within 25 km of Hinkley Point, in which there were 57 cases observed against 57.2 expected

Test	MLR	Pmax	$1/\text{rank}$	$1/\text{distance}$	$\sqrt{1/\text{rank}}$
Power	0.359	0.204	0.421	0.649	0.630
p-value	0.150	0.020	0.108	0.357	0.341

to combine all 13 plants, and there were then sufficient cases to perform a Poisson regression on 1/distance. The resulting estimate of the risk coefficient was positive, but not statistically significant.

37.7.3 Summary of Recommendations

In summary, this is an area of geographical epidemiology where some progress has been made in identifying efficient procedures, perhaps because the problem is essentially one dimensional. Because data sets are usually small, it is an especially important aim to use tests of maximum power, and this criterion seems to be sensitive to the population distribution as well as the precise alternative considered. As far as areal data are concerned, it is recommended that a study should be guided by the following considerations:

1. First and foremost, thought should be given to the patterns of risk that it is desired to detect; these can be expressed in terms of the RRF and may reasonably be supposed to be monotonic decreasing unless special circumstances prevail. The more specifically this can be linked to a biological hypothesis, the more convincing a positive result will be.
2. Next, a circular region of radius R around $\mathcal{S}$ should be chosen and the observed and expected numbers of cases for the areas A_i in R obtained. There is no great advantage for testing purposes in calculating the numbers within fixed distance bands from $\mathcal{S}$. The magnitude of R is important since, if it is much greater than the distance of any conceivable risk, the analysis will inevitably lose power. As a guideline, it would seem sensible to choose the radius R so that the excess relative risk might reasonably be supposed to have declined to half its value at distance $R/2$.
3. A Poisson regression should be used only if the total number of observations is large enough to ensure convergence of the estimation procedure and to provide reliable estimates of the parameters. It is difficult to provide guidelines, but an analysis with fewer than 20 cases in R should be treated with caution. The alternative of a non-parametric test may then be preferable.
4. For a non-parametric test, the first choice to make is between the conditional and the unconditional versions. This will depend largely on the perceived reliability of the expectations and whether it is desired to detect an overall excess in the area as well as spatial pattern.
5. Among tests of either kind, the LRS canonical tests will be reasonably powerful against most monotonic hypotheses, and it is recommended that 1/distance or

$1/\sqrt{\text{distance rank}}$ be used unless the population distribution is very unusual or unless a very non-standard RRF is suspected. In either case, it is recommended that a simulation study be undertaken to determine the most powerful test for the suspected alternatives.

6. The analysis should then proceed with the test identified as best, using simulation to perform a Monte Carlo test unless the expectations are quite large, in which case normal approximations can be used for assessing the significance.

37.8 Conclusions

In this chapter, we have attempted to give a simple but unifying overview of the statistical methods that underlie geographical epidemiology. We have been able to refer to only a small proportion of the very large number of methods that have been proposed for different aspects of the subject. For further reading, we refer to edited volumes by Elliott et al. (1992, 2000), Lawson et al. (1999), and to the *Encyclopedia of Biostatistics* edited by Armitage and Colton (1998), for example the review article by Bithell (1998).

It will be clear that the rational choice of method is not an easy matter. Although the classical theory of statistics provides a number of principles leading to optimal procedures, there are areas of geographical epidemiology where they do not apply. In the first place, they apply essentially to the frequentist paradigm: the increasingly popular Bayesian methods raise essentially new optimality issues that are not easy to resolve. Secondly, many optimality results are asymptotic: when observations are effectively widely distributed throughout two-dimensional space, asymptotic results are less likely to be applicable even in moderately large datasets. Thirdly, many methods are essentially non-parametric and the classical optimality theory applies less directly to these. Lastly, the theoretical results apply mostly to situations where there is a large degree of independence in the structure of the data; they are therefore less applicable to models for the contagious processes needed to model alternatives to the null hypotheses in studies on clustering.

It follows that evaluating the relative merits of different methods has in practice to proceed by largely empirical methods, making extensive use of simulation. This makes appraisal difficult because of the large number of parameters that can be varied in the simulation experiments. It is important that any general principles suggested by the underlying theory are used to direct the empirical investigations, as exemplified, for example, by the discussion of methods for predefined sources of risk in Sect. 37.7. We conclude that geographical epidemiology, despite its practical limitations, can in principle provide useful pointers to the etiology of disease but that the methodology would be much more convincing if we knew more about its behavior in various plausible situations.

References

Abe O (1973) A note on the methodology of Knox's tests of "Time and Space Interaction." Biometrics 29:67–77

Anderson NH, Titterington DM (1995) Some methods for investigating spatial clustering, with epidemiological applications. J R Stat Soc Ser A 160:87–105

Antó JM, Sunyer J (1992) Soya bean as a risk factor for epidemic asthma. In: Elliott P, Cuzick J, English D, Stern R (eds) Geographical environmental epidemiology: methods for small-area studies. Oxford University Press for World Health Organization, Oxford, pp 323–341

Armitage P, Colton T (eds) (1998) Encyclopedia of biostatistics. Wiley, Chichester

Baris YI, Simonato L, Saracci R, Winkelmann R (1992) The epidemic of respiratory cancer associated with erionite fibres in the Cappadocian region of Turkey. In: Elliott P, Cuzick J, English D, Stern R (eds) Geographical environmental epidemiology: methods for small-area studies. Oxford University Press for World Health Organization, Oxford, pp 310–322

Besag J, Newell J (1991) The detection of clusters in rare diseases. J R Stat Soc Ser A 154:143–155

Bithell JF (1990) An application of density estimation to geographical epidemiology. Stat Med 9:691–701

Bithell JF (1995) The choice of test for detecting raised disease risk near a point source. Stat Med 14:2309–2322

Bithell JF (1998) Geographical analysis. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics. Wiley, Chichester, pp 1701–1716

Bithell JF (1999) Disease mapping using the relative risk function estimated from areal data. In: Lawson AB, Biggeri A, Böhning D, Lesaffre E, Viel J-F, Bertollini R (eds) Disease mapping and risk assessment for public health decision making. Wiley, Chichester, pp 247–255

Bithell JF (2003) Selecting a powerful test for detecting risk near point sources. Bulletin of the international statistical institute 54th session. The Institute, Berlin/Germany, pp 97–98

Bithell JF, Dutton SJ, Draper GJ, Neary NM (1994) Distribution of childhood leukaemia and non-Hodgkin's lymphomas near nuclear installations in England and Wales. Br Med J 309(6953):501–505

Bithell JF, Dutton SJ, Neary NM, Vincent TJ (1995) Use of regression methods for control of socio-economic confounding. J Epidemiol Community Health 49(Suppl 2):S15–S19

Clayton D, Bernardinelli L (1992) Bayesian methods for mapping disease risk. In: Elliott P, Cuzick J, English D, Stern R (eds) Geographical environmental epidemiology: methods for small-area studies. Oxford University Press for World Health Organization, Oxford, pp 205–220

Clayton D, Kaldor J (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43:671–682

Cliff AD, Ord JK (1981) Spatial processes: models and applications. Pion, London

COMARE (2005) Tenth report. The incidence of childhood cancer around nuclear installations in Great Britain. HMSO, London

COMARE (2011) Fourteenth report. Further consideration of the incidence of childhood cancer around nuclear power plants in Great Britain. HMSO, London

Cook DG, Pocock SJ (1983) Multiple regression in geographical mortality studies, with allowance for spatially correlated errors. Biometrics 39:361–372

Cox DR, Snell EJ (1984) The analysis of binary data, 2nd edn. Chapman and Hall, London

Cuzick J, Edwards R (1990) Spatial clustering for inhomogeneous populations (with discussion). J R Stat Soc Ser B 52:73–104

David FN, Barton DE (1966) Two space-time interaction tests for epidemicity. Br J Prev Soc Med 20:44–48

Davies TM, Hazelton M (2010) Adaptive kernel estimation of spatial relative risk. Stat Med 29:2423–2437

Dempster AP, Schatzoff M (1965) Expected significance levels as a sensitivity index for test statistics. J Am Stat Assoc 60:420–436

Diggle PJ (2000) Overview of methods for disease mapping and its relationship to cluster detection. In: Elliott P, Wakefield JC, Best NG, Briggs DJ (eds) Spatial epidemiology: methods and applications. Oxford University Press, Oxford, pp 87–103

Diggle PJ, Elliott P (1995) Disease risk near point sources: statistical issues for analyses using individual or spatially aggregated data. J Epidemiol Community Health 49(Suppl 2):S20–S27

Diggle PJ, Rowlingson BS (1994) A conditional approach to point process modelling of elevated risk. J R Stat Soc Ser A 157:433–440

Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman and Hall, New York

Elliott P, Cuzick J, English D, Stern R (eds) (1992) Geographical environmental epidemiology: methods for small-area studies. Oxford University Press for World Health Organization, Oxford, pp 323–341

Elliott P, Wakefield JC, Best NG, Briggs DJ (eds) (2000) Spatial epidemiology: methods and applications. Oxford University Press, Oxford, pp 87–103

Geary RC (1954) The contiguity ratio and statistical mapping. Inc Stat 5:115–145

Gelman A, Price PN (1999) All maps of parameter estimates are misleading. Stat Med 18:3221–3234

Greenland S, Morgenstern H (1989) Ecological bias, confounding, and effect modification. Int J Epidemiol 18(1):269–274

Jacquez GM (1996) A k nearest neighbour test for space-time interaction. Stat Med 15:1935–1949

Kelsall JE, Diggle PJ (1998) Spatial variation in risk: a nonparametric binary regression approach. Appl Stat 47:559–573

Knox EG (1964) The detection of space-time interactions. Appl Stat 13:25–29

Kulldorf M, Hjalmars U (1999) The Knox method and other tests for space-time interaction. Biometrics: 55:544–552

Lawson AB (1993) On the analysis of mortality events associated with a prespecified fixed point. J R Stat Soc Ser A 156:363–377

Lawson AB, Biggeri A, Böhning D, Lesaffre E, Viel J-F, Bertollini R (eds) (1999) Disease mapping and risk assessment for public health decision making. Wiley, Chichester

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman and Hall, London

Moran PAP (1948) The interpretation of statistical maps. J R Stat Soc Ser 10:243–251

Pfeiffer D, Robinson T, Stevenson M, Stevens K, Rogers D, Clements A (2008) Spatial analysis in epidemiology. Oxford University Press, Oxford

Pike MC, Smith PG (1974) A note on a ‘close pairs’ test for space clustering. Br J Prev Soc Med 28:63–64

Potthoff RF, Whittinghill M (1966) Testing for homogeneity II. The Poisson distribution. Biometrika 53:183–190

Rogerson PA (2006) Statistical methods for the detection of spatial clustering in case-control data. Stat Med 25:811–823

Scott DW (1992) Multivariate density estimation. Wiley, London

Silverman BW (1986) Density estimation for statistics and data analysis. Chapman and Hall, London

Snow J (1855) On the mode of communication of cholera, 2nd edn. Churchill, London

Stone RA (1988) Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. Stat Med 7:649–660

Tango T (2010) Statistical methods for disease clustering. Springer, New York

Wakefield JC (2008) Ecological studies revisited. Ann Rev Public Health 29:75–90

Wakefield JC, Best NG, Waller L (2000) Bayesian approaches to disease mapping. In: Elliott P, Wakefield JC, Best NG, Briggs DJ (eds) Spatial epidemiology: methods and applications. Oxford University Press, Oxford, pp 87–103

Walter SD (1993) Assessing spatial patterns in disease rates. Stat Med 12:1885–1894

Statistical Methods in Genetic Epidemiology

38

Heike Bickeböller and Duncan C. Thomas

Contents

H. Bickeböller (✉)
Department of Genetic Epidemiology, University Medical School, Georg-August-University of Göttingen, Göttingen, Germany

D.C. Thomas
Department of Preventive Medicine, Division of Biostatistics, University of Southern California, Keck School of Medicine, Los Angeles, CA, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_62,

38.1 Introduction

Genetic epidemiology combines the scientific disciplines of human genetics, epidemiology, and biostatistics and has close relationships with the fields of medicine, molecular genetics, and molecular epidemiology. The latter traditionally has been concerned more with the study of molecular markers of exposure, susceptibility, and disease (see chapter ▶Molecular Epidemiology of this handbook). The field is also a specialized subdiscipline of biometry and mathematical population genetics with major biometrical contributions to human genetics and the development of statistical methods including segregation, linkage, and association analysis; simulation methods; and computer algorithms. Rather than focusing on cells or molecules (as in molecular genetics) or on individual patients (as in clinical genetics), genetic epidemiology research is conducted using populations or large series of systematically collected families (Khoury et al. 1993).

Genetic epidemiology aims to detect the genetic origin of phenotypic variability in humans (Vogel 2000) and unravel genetic components that contribute to the development or the course of a disease (or more generally a *phenotype*, the observed trait), along with environmental or other risk factors that may modify the effects of genes. Thus, the International Society of Genetic Epidemiology (IGES 2012) describes the field as a marriage between the disciplines of genetics and epidemiology, emphasizing the need to join the fields. Whereas genetics tends to focus on the genotype-phenotype correlation neglecting the environment and epidemiology tends to focus on environmental and demographic factors, a full understanding of the etiology of complex traits can only be achieved by considering both explaining how genes are expressed in the presence of different environmental contexts and how genetic and environmental factors act together in shaping a phenotype.

In contrast to classical risk factor epidemiology, the three main complications in genetic epidemiology are dependencies, large data sets, and the use of indirect evidence. The structure of chromosomes and families or populations leads to major dependencies within the data, thus requiring customized models and tests. Modern technologies can yield millions of genotypes per subject for many thousands of subjects at an affordable cost, and even higher density sequencing platforms are now becoming available, along with a plethora of other data types (e.g., expression, proteomics) and repositories of biological knowledge (*ontologies*). In many studies, the disease-related functionally relevant DNA variant(s) in a gene are not directly observed, and hence the evidence on them is only indirectly given through correlated variant(s).

This chapter is solely devoted to methods dissecting the genotype-phenotype correlation with a binary phenotype (affected/unaffected). It does not specifically cover quantitative phenotypes, although many of the techniques discussed below can be applied to such problems. Section 38.2 presents an overview of major study designs and types of analysis. Section 38.3 introduces the most important genetic models. Sections 38.4–38.6 cover the three major types of analysis, segregation, linkage, and association analysis. Section 38.7 describes recent developments and looks to the future. We offer some general conclusions in Sect. 38.8. Detailed information on diseases used as examples in this chapter may be found in the standard reference of McKusick (1998) or its online version, Online Mendelian Inheritance in Man (OMIM 2012).

38.2 Study Types

Genetic epidemiology investigations are usually triggered by epidemiological studies that demonstrate a positive family history as a risk factor for disease, suggesting the existence of genetic or shared environmental factors. Often the goal of initial studies is to estimate the *relative risk for relatives of affected individuals* compared to the general population, such as λ_S in the case of siblings of affected individuals, in order to support a genetic hypothesis.

To further investigate familial aggregation, a *segregation analysis* may be carried out in pedigrees. This aims to determine whether a *major gene* is influencing a given phenotype in these families and if so to estimate the parameters of the underlying genetic model. All methods for segregation analysis are based on probability calculations for the observed phenotypes conditional on hypothetical genetic model parameters and on family structure, that is *genealogies*. Parameter estimation is often based on likelihood-ratio tests in order to select the most plausible model nested within a hypothetical general model. Sometimes family studies are also used solely (twin studies) or jointly with a major gene to estimate the *heritability* h^2 of a trait, that is, the proportion of the variance explained by (additional) genetic components. Hence, λ_S or h^2 are often used to indicate the genetic basis of a phenotype in a population (or enriched families) before marker studies are performed.

The primary cause of a *monogenic disease* such as cystic fibrosis is a mutation within a single gene that segregates according to Mendelian laws (see below). The predisposing variants (the alleles carrying the risk) of this major gene are usually rare in the population. For *complex* or *multifactorial diseases*, there may still be Mendelian subforms such as breast cancer caused by the major gene *BRCA1*. For rare monogenic diseases and rare Mendelian subforms of complex diseases, segregation analysis and subsequent further analyses perform well. However, complex diseases in general require more sophisticated methods of analysis. For example, in Alzheimer's disease, there are at least three major genes and several susceptibility genes conferring moderate risk (*oligogenes*). Oligogenes can have relatively common alleles carrying the risk. *Polygenic* effects at many loci across the whole genome, each with a minor effect, may contribute to disease.

If there is evidence for the existence of genetic factors contributing to a disease, the next step is to identify susceptibility genes in order to quantify the genetic influence and to understand the underlying genetic model and pathway to the phenotype. To this end, measures of correlation between a *genetic marker* and the (unknown) *disease locus* are used. A genetic marker is a DNA segment for which the chromosomal localization is known and multiple alleles can be determined. In general, methods assume Mendelian segregation of the marker (see Sect. 38.3.2).

Frequently used markers are multiallelic *microsatellites* and biallelic *single nucleotide polymorphisms (SNPs)*. A marker is termed a *polymorphism* if the frequency of the most common variant is less than 99%.

Two types of correlation between a genetic marker and the susceptibility locus are used:

- *Linkage (cosegregation at the family level)*: Linkage is present if the transmissions of DNA at marker and disease susceptibility loci from a parent to a child are not independent. Relatives with a similar disease status (e.g., both affected) are then more similar at a marker close to the disease susceptibility locus than expected under independence.
- *Linkage disequilibrium (LD, association at the population level)*: LD is present if in a gamete the joint probability for a specific marker allele and a specific disease allele differs from the product of individual probabilities. In affected individuals, certain marker alleles will then be more or less frequent than in randomly selected individuals from the population.

Linkage analysis in families is based on linkage; *association analysis* in populations or families uses linkage disequilibrium. Some designs and corresponding statistical methods are capable of integrating both types of information into the analysis.

For the analysis of complex diseases with genetic markers, we can distinguish two major approaches: A *candidate gene investigation* focuses on genes (or genomic regions) whose function in the pathway to the phenotype is thought to be known. A prominent example of a candidate gene system is the HLA (human leukocyte antigen) complex on chromosome 6. HLA is involved in immune resistance and is thus a natural candidate gene region for all autoimmune diseases. The genotypes of the relevant functional component of the candidate genes are not always observed. In this case, we use the information on genetic markers that lie in or in close proximity to the candidate gene in question. In contrast, a *genome scan* – a systematic coarse grid search of the whole genome with genetic markers – aims to localize one or more regions harboring susceptibility genes. A typical scan might investigate approximately 350–700 microsatellites with an average distance of 5–10 cM (centiMorgan, see Sect. 38.3.3) or 500,000 or more SNPs along the genome, depending on the type of genetic information and design used (linkage or association, respectively).

38.3 Genetic Models

38.3.1 Terminology

The *genome* is the complete collection of an individual's genetic material present in every cell, consisting of *chromosomes* (long DNA strands). A *gene* is a piece of a chromosome coding for a function that can be seen as the heritable unit. The *locus* is the position of a piece along the chromosome. The locus might denote the position of, for example, a gene, a gene complex, or a marker. The different variants of a gene are called *alleles*.

The human genome is *diploid*, that is, chromosomes are paired *(homologous chromosomes)* with the exception of the sex chromosomes in males. Each human somatic cell contains 22 *autosomal* pairs and 1 pair of sex chromosomes. In a pair, the autosomal chromosomes contain the same gene with possibly different alleles at the same location. During *meiosis*, a diploid chromosome set is reduced to a *haploid* chromosome set of a germ cell, the *gamete*.

A pair of an individual's alleles at a locus is called a g*enotype*. If the alleles are identical, the individual is called *homozygous* at the locus, otherwise *heterozygous*. Two copies of a gene are called *identical by descent (IBD)* if both copies are identical and are copies of the same gene in a common ancestor. An individual is *homozygous by descent (HBD)* when its gene pair is IBD. When considering several loci simultaneously, the multilocus alleles inherited from the same parent constitute a *haplotype*.

38.3.2 Mendelian Single-Locus Model

Mendelian segregation (Mendel 1865) is the simplest and most commonly used model for the *mode of inheritance* for a single locus. An individual randomly and independently inherits one allele from father and mother, respectively. All segregation events from parents to offspring are independent. This implies that copies of some alleles are frequently present in offspring and other alleles are lost in subsequent generations, hence leading to random changes in population allele frequencies over time (*genetic drift*).

Consider the phenotype affected/unaffected by a certain disease. Let S denote a susceptibility gene with n alleles $S_1,\ S_2,\ldots,S_n$. The distribution of allele frequencies in the population is denoted by $p_r = P(S_r)$ $(r = 1,\ldots,n)$. Under Hardy-Weinberg equilibrium (HWE) (Hardy 1908; Weinberg 1980), the (unordered) genotype frequencies are given by

$$\begin{array}{ll} P(S_r S_s) = p_r p_s = p_r^2 & \text{for } r = s \\ P(S_r S_s) = 2 p_r p_s & \text{for } r \neq s \end{array}.$$

These frequencies follow from independence of the corresponding allele frequencies, combining two ordered genotypes for heterozygotes. Its maintenance in a population can be derived by applying Mendelian segregation to each possible parental mating type, assuming random mating (Khoury et al. 1993).

Penetrance describes the relation between genotype and phenotype. It is the conditional probability that an individual will be affected given its genotype: $f_{rs} = P(\text{affected} \mid S_r S_s)$.

Classical monogenic diseases are those caused by a single major gene, for which the penetrances take only the values 0 or 1. Often a locus S is assumed to be *biallelic*, that is, to have only two different alleles. Let S_1 denote the "susceptibility" allele (mutation) and S_2 the "normal" allele (wild type). For a classical *dominant* *disease,* all carriers of the susceptibility allele are affected ($f_{11} = f_{12} = f_{21} = 1$, $f_{22} = 0$); for a classical *recessive disease*, only homozygous carriers are affected ($f_{11} = 1$, $f_{12} = f_{21} = f_{22} = 0$).

Many classical hereditary diseases follow a Mendelian mode of inheritance. Often, the prevalence is below 1 in 1,000 live births. Examples are cystic fibrosis (autosomal recessive gene, cystic fibrosis transmembrane regulator (*CFTR*)) and Huntington chorea (autosomal dominant gene, *huntingtin* (HTT)). Both diseases can be caused by any of many different mutations. However, the assumption of a gene with a normal and a susceptibility allele (group) worked well in identifying these genes as disease causes, even though the true inheritance is much more complicated. The aim in statistical genetics is not to specify the correct and complete model but to address the scientific question adequately with a parsimonious mathematical model. If this model is too simple, then more complex or new biologically motivated models need to be considered.

The genotype-phenotype relation is complicated for many diseases. Individuals with a susceptible genotype can stay unaffected (*incomplete penetrance*) and those with a non-susceptible genotype can become affected (*phenocopies*). The penetrances at a specific gene locus may be different for different alleles and may depend on age, sex, environmental exposures, or other factors. For a general single locus with susceptibility allele S_1, we generally assume that $1 \geq f_{11} \geq f_{12} = f_{21} \geq f_{22} \geq 0$ and specifically for a recessive mode of inheritance that $f_{12} = f_{21} = f_{22}$ or for a dominant one that $f_{11} = f_{12} = f_{21}$. It is usually assumed that the parental origin of an allele has no influence on a disease, that is, $f_{12} = f_{21}$, although there is a growing literature on imprinting and parent-of-origin effects that violate this assumption (Zhou et al. 2010; Ainsworth et al. 2011).

38.3.3 Linkage

For the joint inheritance at two loci, independent Mendelian segregation does not generally hold, owing to crossover events and recombinations. *Gametes* (comprising one allele from each pair of chromosomes) are formed during meiosis, when homologous chromosomes are arranged next to each other and partly overlap. A chromosome breakage and a *crossover* – an exchange between homologous chromosomal segments – can occur. A *recombination* between loci A and B occurs when a gamete has a haplotype comprising a combination of alleles different from that on the same grandparental chromosome due to crossovers between the loci.

Consider the formation of gametes during meiosis displayed in Fig. 38.1. Between distant loci A and B (see Fig. 38.1a), a crossover is likely to result in a recombination of the haplotypes A_1B_1 and A_2B_2 to give the new haplotypes A_1B_2 and A_2B_1. If the two loci A and B are very close (see Fig. 38.1b), this is very unlikely. *Map distance* is defined as the expected number of crossovers between two loci (Haldane 1919). In expectation, the number of crossovers is roughly proportional to physical distance between two loci, so this distance measure

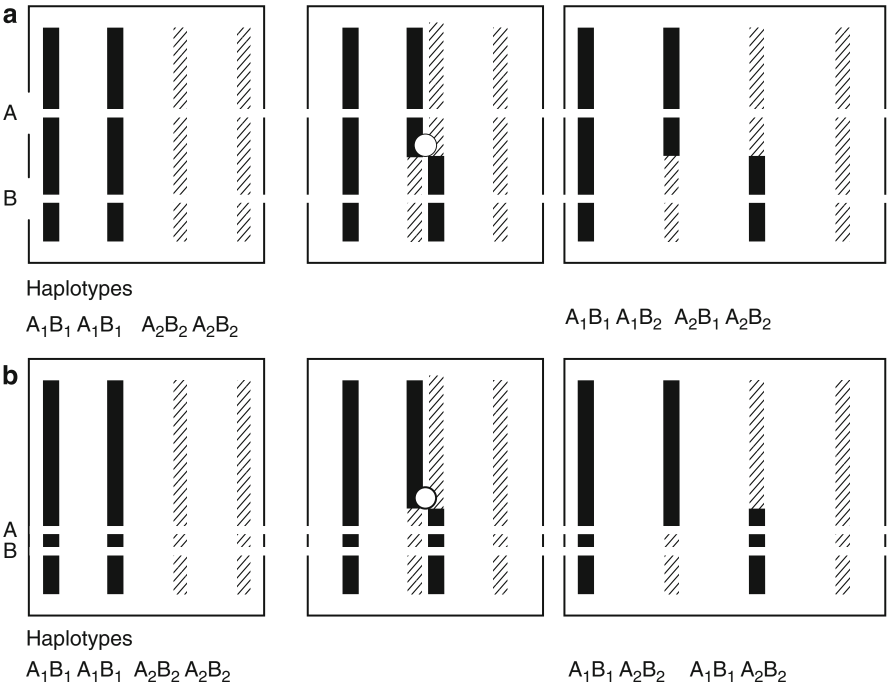


Fig. 38.1 Formation of gametes during meiosis from one parental pair of chromosomes with a single crossover. *Left*: parental chromosome pair, *middle*: crossover event (crossover point denoted by the *circle*), *right*: gametes for offspring formation. At the two loci A and B, the parent is double heterozygous A_1A_2 and B_1B_2. (**a**) The crossover occurred between locus A and B. The two middle gametes show recombination. (**b**) The crossover occurred above locus A and B so that the gametes do not show recombination

is additive, that is, for three (ordered) loci A, B, and C, the map distance between A and C is the sum of the distances between A and B and between B and C. The map unit is called a Morgan (M), named after T.H. Morgan, a Nobel prize winning geneticist (1866–1945), who discovered the importance of chromosomes for the inheritance process. The human genome contains approximately 3.3 billion base pairs with a total length of approximately 33 M, so as a rough guide, 1 centimorgan ($\text{cM} = 0.01\,\text{M}$) corresponds to one million base pairs in the physical map.

By genotyping, it is possible to observe recombinations between two loci, but crossovers are not directly observable. Figure 38.1a shows recombination due to a single crossover. For a double crossover (two chromosomal exchanges between loci), no recombination would be observed. This holds true for any even number of crossovers. Thus, a *recombination* is defined as an uneven number of crossovers between a pair of loci. The *recombination rate* θ – the ratio of the number of recombinant gametes to the total number of gametes formed – is used as a measure

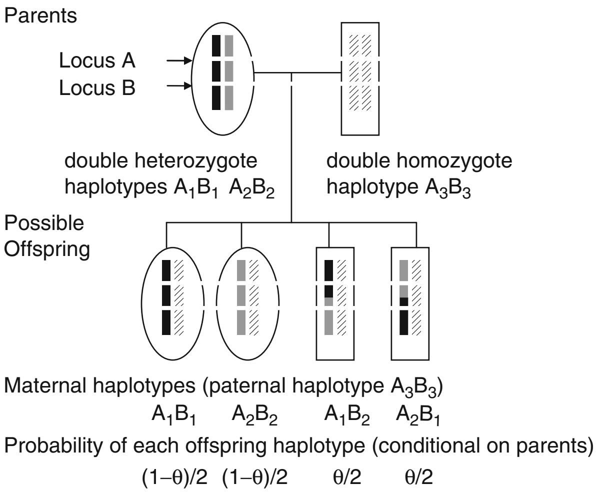

Parents
Locus A
Locus B
double heterozygote haplotypes A_1B_1 A_2B_2
double homozygote haplotype A_3B_3
Possible Offspring
Maternal haplotypes (paternal haplotype A_3B_3)
A_1B_1 A_2B_2 A_1B_2 A_2B_1
Probability of each offspring haplotype (conditional on parents)
(1−θ)/2 (1−θ)/2 θ/2 θ/2

Fig. 38.2 Formation of recombinant and non-recombinant haplotypes by meiosis

of *genetic distance* between two loci. If loci are on different chromosomes or far away on the same chromosome, they segregate independently. By definition, there is *linkage* between the loci if $0 \leq \theta < 0.5$, and no linkage if $\theta = 0.5$. The closer loci are to each other, the less likely there will be crossovers and hence a recombination. Complete linkage (complete cosegregation) implies no recombination, and thus $\theta = 0$.

In Fig. 38.2, a double heterozygous parent with haplotypes A_1B_1 and A_2B_2 and a double homozygous parent with haplotype A_3B_3 are considered. For the double heterozygous parent, a meiosis can create the non-recombinant haplotypes A_1B_1 and A_2B_2 or the recombinant haplotypes A_1B_2 and A_2B_1. In order to determine recombination, a parent homozygous even at one locus is not informative. Given that recombination is present, each of the two recombinant haplotypes occurs with probability 0.5. Given that no recombination is present, each of the two non-recombinant haplotypes occurs with probability 0.5. For $\theta = 0.5$, there is independent segregation so that all four possible haplotypes are equally likely.

If three ordered close loci A, B, and C are considered, $\theta_{AC} \approx \theta_{AB} + \theta_{BC}$. In contrast to the map distance in Morgans, recombination distances are not additive. *Mapping functions* provide a translation of recombination distances into map distance in Morgans. In the majority of chromosomal regions, recombination rates for women are higher than for men.

The potential informativeness of a single marker chosen from an existing marker map (without consideration of the disease locus) is determined by its genetic variability, that is, allele distribution. A measure of marker informativitiy is the *heterozygosity* H, defined as follows (Weiss 1993; Ott 1999):

$$H = \sum_{r \neq s}^{n} p_r p_s .$$

38.3.4 Linkage Disequilibrium

Linkage and linkage disequilibrium (LD) are different concepts. As linkage describes the coinheritance at two loci, it can only be observed in families, and it is independent of the specific alleles. LD describes the relation between alleles at two loci in a population.

Let S denote a locus with n alleles $S_1, S_2, \ldots, S_n$ and allele frequencies $p_r = P(S_r)$ and M a locus with m alleles $M_1, M_2, \ldots, M_m$ and allele frequencies $q_i = P(M_i)$. A common measure of LD is the haplotype probability minus its expectation under no association. For two biallelic loci, it is denoted by D or δ. For multiallelic markers, the parameter δ_{ir} is often used to define the linkage disequilibrium between M_i and S_r as

$$\delta_{ir} = P(M_i, S_r) - P(S_r)P(M_i), \qquad i = 1, \ldots, m;\ r = 1, \ldots, n.$$

LD is present if $\delta_{ir} \neq 0$ for any pair of alleles M_i and S_r. Under LD, the allele distribution at locus M is dependent of the S allele present. Often used measures of LD are D' (Devlin et al. 1996), defined as δ divided by its theoretical maximum for the observed allele frequencies, that is, a rescaling of δ to range between 0 and 1, and R^2, the square of the correlation coefficient.

LD can arise in several different ways (Suarez and Hampe 1994). At linked loci, complete LD can be caused by a recent mutation at one locus. However, disequilibrium is also possible without linkage between the loci (the term *gametic disequilibrium* is preferable in this case). One important mechanism for the development of disequilibrium at unlinked loci, even on different chromosomes, is *population stratification*. For example, through immigration or non-random mating (e.g., by religion or social status), populations may admix with different allele distributions in the populations.

Under random mating, LD decays over generations g according to $\delta_g = (1 - \theta)^g \delta_0$, where δ_0 is the initial LD at generation 0 (Maynard Smith 1989). Thus, whatever the origin of LD, in the presence of tight linkage, it can stay strong during many generations. Without tight linkage, LD will degrade rapidly. Thus, LD provides indirect evidence for linkage.

38.4 Segregation Analysis

The aim of *segregation analysis* is to test for the existence of a major gene influencing a phenotype and to estimate its mode of inheritance. The pattern of inheritance may be investigated in a few large families or in many small families.

Consider a Mendelian single-locus model for a major gene with susceptibility allele S_1 and normal allele S_2. In the classical Mendelian disease model, the penetrances $P(\text{affected} \mid \text{genotype})$ are only 0 or 1, so the genotype directly translates to a phenotype, and the families segregating the S_1 display characteristic disease patterns.

The simplest segregation tests (see, e.g., Sham (1998)) are based on *segregation ratios*, the proportion of affecteds among offspring of particular parental mating types. For illustration, consider a rare autosomal dominant disease and matings between an affected and an unaffected individual. These will usually be $S_1S_2 \times S_2S_2$ *matings*. Let r be the observed number of affecteds among n offspring and q the probability for a child to be affected. Then the segregation ratio is the unknown parameter q of a binomial distribution with sample size n. If the null hypothesis $q = 0.5$ is not rejected, it may be concluded that the data are compatible with an autosomal dominant disease pattern.

For each of six possible mating types ($S_1S_1 \times S_1S_1, S_1S_1 \times S_1S_2, S_1S_1 \times S_2S_2, S_1S_2 \times S_1S_2, S_1S_2 \times S_2S_2, S_2S_2 \times S_2S_2$), the distribution of genotypes and phenotypes in the offspring is determined by various genetic models. However, families are often recruited non-randomly according to particular ascertainment criteria, such as enrichment for disease, yielding an oversampling for particular parental genotypes, so for a valid test, the probability distribution needs to be corrected for this *ascertainment bias*. For example, if all families with at least one affected offspring are recruited (*truncate ascertainment*), the distribution of the number of affected offspring can be corrected for ascertainment by considering a truncated binomial distribution conditioning on $r \geq 1$ per family. If instead each case has an equal probability of being ascertained (*single ascertainment*), then multiple case families are represented proportional to the number affected and simply excluding the proband from the analysis may suffice. In practice, ascertainment schemes may be complex or unsystematic (Elston 1995), and misspecification of ascertainment might cause serious bias in the estimation of genetic parameters (see, e.g., Shute and Ewens (1988)).

For an *extended pedigree* with N individuals, a numerical procedure is needed. Let L denote the likelihood for the observed vector of phenotypes $Y = (Y_1, \ldots, Y_N)$, given a genetic model and the pedigree structure. L can be calculated by summing over all possible genotype vectors $G = (G_1, \ldots, G_N)_i$, $i = 1, \ldots, N$, in a given family, a particular one denoted by $g = (g_1, \ldots, g_N)$. We assume that the phenotype Y_i only depends on genotype G_i of that individual i. Thus, we get

$$L(Y) = \sum_{g_1} \sum_{g_2} \cdots \sum_{g_N} P(Y|G = g)P(G = g).$$

The Elston-Stuart algorithm (Elston and Stewart 1971) provides an efficient recursive formula

$$L(Y) = \sum_{g_1} \sum_{g_2} \cdots \sum_{g_N} \prod_{j=1}^{N} P(Y_j|g_j) \prod_{k=1}^{N_1} P(g_k) \prod_{m=1}^{N_2} \tau(g_m|g_{m1}g_{m2})$$

where N_1 denotes the number of *founders* (individuals without specified parents in the pedigree, i.e., members of the oldest generation and married-in spouses) and

N_2 the number of *non-founders*. The parameters of the genetic model are (1) the genotype distribution $P(g_k)$, $k = 1, \ldots, N_1$, for the founders in the population for which HWE is usually assumed, (2) the transmission probabilities from parents m_1 and m_2 to offspring $m, \tau(g_m|g_{m_1}, g_{m_2})$, according to Mendelian segregation $(\tau(S_1|S_1S_1) = 1$; $\tau(S_1|S_1S_2) = 0.5$; $\tau(S_1|S_2S_2) = 0)$, and (3) the penetrances $P(Y_j|g_j)$ relating genotypes to disease. This recursive formula works well on a *simple pedigree* of arbitrary size. As several unrelated pedigrees are independent from each other, their likelihoods can be multiplied to yield the total likelihood of a sample of pedigrees. Computations on *complex pedigrees* with marriage chains and inbreeding loops (such as consanguineous marriages) are often only possible with approximation methods.

Segregation analysis works well for monogenic diseases. Due to the unclear genotype-phenotype relationship, they are much more difficult in complex diseases. Several genetic and non-genetic factors, such as age, sex, and exposure factors, may have an influence on disease. Genetic heterogeneity may be caused by different alleles of the same gene or different genes or modifying factors that lead to different phenotypes segregating within a family (Evans and Harris 1992). In addition, there are further types of genetic heterogeneity such as genomic imprinting. In the presence of heterogeneity, considering homogeneous subgroups (defined by, say, severity, age of onset, family history, ethnicity) can lead to a clearer genotype-phenotype relation and thus to identify a possibly Mendelian subform of the disease.

An example of a highly successful segregation analysis for a complex disease is breast cancer (Newman et al. 1988). The families were ascertained through a population-based cancer registry. The ascertainment criteria for index cases were women with breast cancer, Caucasian, diagnosed before the age of 55 during a specified period with a histologically confirmed primary tumor. There was no selection on family history. Thousand five hundred and seventy-nine nuclear families were recruited, along with one large extended pedigree. Complex segregation analysis was applied using the program POINTER (Lalouel et al. 1983). It models an underlying unobserved quantitative trait called *liability* as a mixture of three normal distributions with different means for each of the genotypes, allowing for polygenes and an environmental component, with disease corresponding to the liability exceeding a certain threshold (Morton and MacLean 1974). For a predisposing genotype, the mean liability is shifted compared to the mean for non-disposing genotypes such that more individuals will exceed the threshold. Evaluation by direct modeling of the transmission probabilities allows the identification of the major factor as a major Mendelian gene. Population-based liability classes were taken from cumulative incidence rates estimated from a large epidemiological study in the region.

Segregation analysis is based on likelihood-ratio tests comparing different models. First, one investigates the consistency of the data with a major gene model; next, one considers which mode of inheritance fits the data best. To avoid false-positive results, the likelihood for Mendelian transmission probabilities can be compared against more general models corresponding to environmental or cultural transmission. For breast cancer, an autosomal dominant major gene provided the best fit, although the general single-locus model with three penetrances resulted in a comparable fit, and the non-Mendelian transmission models were strongly rejected.

Segregation analysis for a disease without a single major gene but only a few oligogenes may not be particularly rewarding. Many genetic marker studies are nowadays carried out without specification of the genetic model, but it is still worth establishing that a disease has some genetic basis by estimating the heritability h^2 before embarking extensive marker studies.

38.5 Linkage Analysis

In *linkage analysis*, the cosegregation between marker and disease within families is investigated to find evidence for linkage and often to estimate the recombination rate θ. The classical *lod score method* (Morton 1955) is a test for linkage between a susceptibility gene and a marker (null hypothesis H_0: $\theta = 0.5$ vs. alternative H_1: $\theta < 0.5$) under a parametric model for the genetic effect, allowing estimation of θ. For a detailed description, see Ott (1999). Let $L(\theta)$ denote the likelihood for the observed phenotypes at a particular value for θ conditional on the (assumed) model, the marker allele distribution, and the given pedigrees. In the usual notation, the underlying conditioning is sometimes left out. The lod score function ("log odds") is the log likelihood ratio

$$Z(\theta) = \mathrm{LOD}\,(\theta) = \log_{10} \frac{L\,(\theta)}{L\,(0.5)}$$

as a function of θ. $Z(\theta)$ compares the likelihood under linkage with recombination rate θ with the likelihood under no linkage, that is, $\theta = 0.5$. $Z(\theta)$ will be maximized over all possible values for θ, that is, $0 \leq \theta \leq 0.5$, to yield $Z_{\max}$. Values of $Z_{\max} > 3$ are taken as evidence for linkage. The recombination rate is estimated by $\theta_{\max}$, the θ-value corresponding to $Z_{\max}$. If $Z_{\max} < -2$, linkage can be excluded. The limits 3 and -2 are based on a sequential Wald test, such that the posterior probability for linkage when rejecting H_0 is 95% for a single alternative θ. As logarithms of base 10 are used, the limits correspond to stopping limits of 1,000 and 0.01 in the sequential testing procedure yielded by setting the probabilities of types I and II errors at $\alpha = 0.001$ and $\beta = 0.01$ (Morton 1955).

The likelihood $L(\theta)$ for linkage between two loci A and B for a sibship can be derived easily when the genotypes at both loci are directly observable. Let the mothers' genotypes be A_1A_2, B_1B_2 and the fathers' A_1A_1, B_1B_1; here, only the double heterozygous mother is informative. Without knowing the maternal grandparent's genotypes, one cannot determine the *phase*, that is, whether the mother's haplotypes are A_1B_1 and A_2B_2 (phase I) or A_1B_2 and A_2B_1 (phase II). If the phase were known, $L(\theta)$ would be given by a binomial distribution with

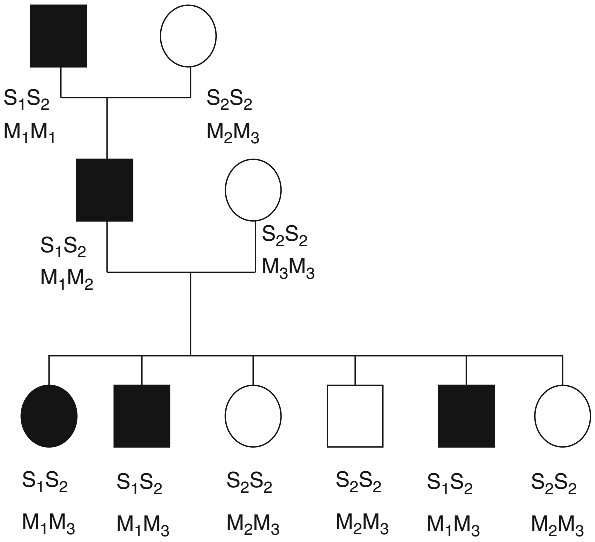


Fig. 38.3 Pedigree with a sibship of size six with marker information and with genotype information concerning the susceptibility locus, owing to the clear-cut rare autosomal dominant mode of inheritance (individuals: square = male, circle = female, black = affected, white = unaffected)

parameters n and θ, where n is the number of informative meioses. For an unknown phase, consider first phase I: Let n_x and n_y denote the number of meioses from the mother to the n children, of which n_x are non-recombinants (A_1B_1 or A_2B_2) and n_y are recombinants (A_1B_2 or A_2B_1). Under phase II, let n_x and n_y denote instead the number of recombinants and non-recombinants. Assuming LD phases I and II are both equally likely, then

$$L(\theta) = \binom{n_x + n_y}{n_x} \left[\frac{1}{2} (1 - \theta)^{n_x} \theta^{n_y} + \frac{1}{2} \theta (1 - \theta)^{n_x} \theta^{n_y} \right].$$

For the sibship in Fig. 38.3, let us now determine the likelihood $L(\theta)$, the lod score function $Z(\theta)$, $Z_{\max}$, and $\theta_{\max}$. Assume an autosomal dominant gene S with a rare susceptibility allele S_1 and a normal allele S_2. Thus, the affected father and all affected siblings have genotype S_1S_2. The marker M has alleles M_1, M_2, and M_3. The mother is homozygous and uninformative for linkage. She will not be considered further.

As a result of the genotyped grandparents, the father's haplotypes are known: S_1M_1 and S_2M_2. Thus, the phase is known and the likelihood is

$$L(\theta) = \binom{6}{0} (1 - \theta)^6 \theta^0 = (1 - \theta)^6 .$$

The lod score function is

$$\begin{aligned} Z(\theta) = \mathrm{LOD}\,(\theta) = \log_{10}\tfrac{L(\theta)}{L(0.5)} &= \log_{10}\tfrac{(1-\theta)^6}{(0.5)^6} \\ &= 6\log_{10}(1-\theta) + 6\log_{10}2 \\ &= 6\log_{10}(1-\theta) + C \end{aligned}$$

where C denotes a constant independent of θ. The maximum of the lod score function is $Z_{\max} = 1.8$ for $\theta_{\max} = 0$. This corresponds to complete linkage as supported by no observed recombinations.

Missing information on grandparental genotypes in Fig. 38.3 results in an unknown phase. Then the lod score function would be

$$\begin{aligned} Z(\theta) = \mathrm{LOD}\,(\theta) &= \log_{10}\tfrac{L(\theta)}{L(0.5)} = \log_{10}\tfrac{0.5\theta^6+0.5(1-\theta)^6}{(0.5)^6} \\ &= \log_{10}\left(\theta^6 + (1-\theta)^6\right) + 5\log_{10}2. \end{aligned}$$

In this case, the maximum of the lod score function is $Z_{\max} = 1.5$ for $\theta_{\max} = 0$. Due to the uncertain phase, the maximum lod score is reduced. However, the estimate for the recombination rate stays at $\theta = 0$.

In Fig. 38.3, assume now that the second affected child has the genotype M_2M_3 (and the genotype S_1S_2). With the phase as indicated in the figure, one recombination needs to be taken into account now. Thus,

$$\begin{aligned} Z(\theta) = \mathrm{LOD}\,(\theta) &= \log_{10}\tfrac{L(\theta)}{L(0.5)} = \log_{10}\tfrac{6\theta(1-\theta)^5}{6(0.5)^6} \\ &= \log_{10}\theta + 5\log_{10}(1-\theta) + 6\log_{10}2. \end{aligned}$$

With one recombination, the maximum of the lod score function is $Z_{\max} = 0.63$ for $\theta_{\max} = 1/6 = 0.17$. Now linkage is estimated as incomplete, and $Z_{\max}$ is markedly reduced.

If in Fig. 38.3 the genotypes of the father and his parents are unknown, the father's genotype can be inferred as either M_1M_1 or M_1M_2. If HWE can be assumed, the likelihood of the recombination rate $L(\theta)$ can be calculated as a function of the marker allele frequencies in offspring. A detailed calculation will show that in this case, a rare marker allele M_1 will result in a high lod score and a more common marker allele M_1 will result in a lower lod score.

For a larger pedigree, $L(\theta)$ can be computed by the Elston-Stuart algorithm (Elston and Stewart 1971) described earlier, where now the g_j are the joint genotypes formed by the two loci S and M and θ is part of the transmission probabilities for the formation of gametes as recombinants or non-recombinants. Across families, the segregation process is independent, so $L(\theta)$ is simply the product of the individual family-specific likelihoods.

The lod score method has been very successful in localizing major genes, such as *BRCA1* for breast cancer (Hall et al. 1990), which was facilitated by focusing on early-onset families for which the genetic relative risk (*RR*) is strong and by having available a good segregation model. However, in complex diseases, the mode of inheritance is usually unclear, leading to false-positive and false-negative results as well as biased estimation of θ (Risch 1991; Lander and Schork 1994; Terwilliger and Ott 1994; Ott 1999). Joint segregation and linkage analysis often leads to biased parameter estimation, particularly if families are not systematically ascertained or the segregation model is misspecified. In MOD-score analysis (Risch 1984; Clerget-Darpoux et al. 1986), the LOD score is maximized over θ and the parameters of a biallelic single-locus model, that is, allele frequency and three penetrances. *Non-parametric methods* or *model-free methods* have been developed to avoid assumptions about the underlying genetic model. Their aim is to provide evidence for linkage without specifying the parameters of the underlying mode of inheritance and without estimating the recombination rate (Lander and Schork 1994; Elston 1998). They are often based on the *identity-by-descent (IBD)* status (Penrose 1953). For example, the IBD status of a patient and one of his/her siblings can take on the values 0, 1, or 2, according to the number of marker alleles that have been transmitted to both siblings from exactly the same grandparental copy of a parent's gene and are thus identical.

Allele sharing methods test whether relatives with a similar disease status (e.g., both affected) are more frequently similar in IBD at the marker than expected in the absence of linkage. In the *affected-sib-pair (ASP) method* (Day and Simons 1976), the observed counts of n ASPs with 0, 1, or 2 marker alleles IBD are compared with the expected ones assuming no linkage ($0.25n$, $0.5n$, or $0.25n$) using a χ^2-goodness-of-fit-test.

The literature on IBD methods is extensive and more powerful methods have been developed (e.g., Holmans 1993; Whittemore and Tu 1998). To determine IBD unambiguously, the marker must be sufficiently polymorphic, and the parents must be genotyped, or neighboring loci must be genotyped to yield sufficient information on the grandpaternal inheritance of the alleles. Often, IBD needs to be estimated. Sometimes, the *identity-by-state* (*IBS*) status (the number of identical marker alleles without considering ancestry) is used instead. In the Lander-Green algorithm for multipoint linkage analysis (Lander and Green 1987), the inheritance vector and thus the IBD status at a particular locus can be determined much more precisely even when some parents are not genotyped. The algorithm calculates the inheritance vector using a hidden Markov model walking from marker to marker, where pedigree size that a software can handle is limited by the length of this vector determined by the number of founders and non-founders in the pedigree (Lander and Green 1987; Kruglyak et al. 1996).

38.6 Association Analysis

The aim of *association studies* is to provide evidence for association or linkage disequilibrium in a population. LD results in an association between marker alleles and alleles of a susceptibility gene, such that certain marker alleles will be present more often in affected individuals than in a random sample of individuals from the population.

In classic *case-control studies,* marker allele frequencies or genotype frequencies in a group of unrelated affected individuals are compared to those in a group of unrelated unaffected individuals. Numerous associations have been identified with case-control studies, for example, associations of autoimmune diseases (e.g., diabetes, multiple sclerosis) with the HLA system or of apolipoprotein E (*APOE*) allele ε4 with Alzheimer's disease (Corder et al. 1993). The *APOE* ε4 allele frequency is approximately 35% in Alzheimer's patients, but only 15% in the older population not suffering from dementia. If a positively associated marker allele is frequent in a population, such as *APOE* ε4, then it is by itself not a good predictor for disease status, and the proportion of homozygotes for the allele is high. Linkage analysis methods are in general not very powerful in this situation.

Besides the usual limitations of classical case-control studies in epidemiology (cf. chapter ▸Case-Control Studies of this handbook), case-control studies to investigate linkage disequilibrium in genetic epidemiology must take a particular form of confounding known as population stratification into account. Population stratification denotes the presence of different ancestry populations, that is, discrete subpopulations or admixture of populations. If individuals are descended from populations with different allele frequencies and this is not taken into account, then spurious associations can be induced. To avoid this confounding, cases and controls must originate from the same homogeneous (including ethnically homogeneous) source population, or an appropriate design and analysis strategy needs to be employed.

If an association is found that is not considered spurious, this may have two causes (Lander and Schork 1994):

- The disease-associated allele is the susceptibility allele itself. If so, this association is expected to occur in all populations harboring this allele.
- The associated allele is in linkage disequilibrium with the susceptibility allele at the disease locus. If this is the case, then different associations can occur in different populations due to different haplotype frequencies at the two loci.

In the first case, marker and disease loci are identical, so $\theta = 0$, and LD is complete. In the second case, marker and disease locus are in general very close to each other. For this reason, association studies are highly valuable for the investigation of candidate genes.

As mentioned above, uncontrolled population stratification can result in spurious associations. For case-control studies, there are methods for taking the existence of subpopulations into account during statistical testing. All methods require many markers along the genome to be genotyped. In the *genomic control* method (Devlin and Roeder 1999), a variance inflation factor is used to adjust the test statistic, taking into account correlations between individuals in subpopulations. The *structured*

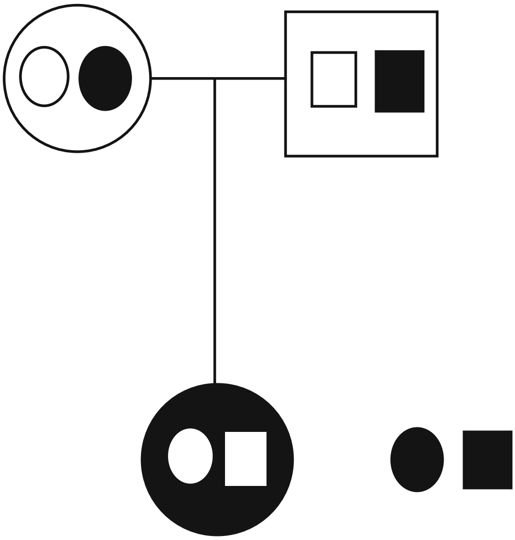


Fig. 38.4 Nuclear family with one affected child. Alleles transmitted from the parents to the affected child are denoted in *white*. Alleles not transmitted from the parents to the affected child are denoted in *black* (individuals: square = male, circle = female, black = affected, white = unaffected, inside: alleles)

association method (Pritchard et al. 2000) estimates the population structure and either assigns individuals to the most likely subgroup or better, estimates the proportion of each individuals' genome derived from each subgroup. Association is subsequently tested within subgroups or adjusted for ancestral source proportions. Both methods typically use panels of hundreds of "ancestrally informative" markers. *Principal component analysis* (Price et al. 2006) generally uses all the markers from a genome-wide scan and adjusts the association of any particular marker for the first few dozen principal components.

Family-based association studies avoid bias due to inadequate controls and population stratification by design. The concept of *internal controls* was first proposed by Falk and Rubinstein (1987). For the original design, nuclear families with at least one affected child are recruited, and the two parental alleles not transmitted to the affected child are used as internal controls (Fig. 38.4). With this design, information on both linkage and association between a marker and the susceptibility gene is used.

For a biallelic marker, the data resulting from this study design can be presented in various ways as 2 × 2 contingency tables and analyzed with standard statistical tests to investigate whether certain alleles are transmitted from the parents to an affected child more often than not (Terwilliger and Ott 1992; Schaid and Sommer 1994). Although in principle, all these procedures test for association (H_0: $\delta = 0$ vs. H_1: $\delta \neq 0$) and most for linkage as well, the most appropriate test respecting the matched nature of the transmitted and non-transmitted allele data is the McNemar test, which in this context is known as the *Transmission/Disequilibrium Test* (*TDT*)

Table 38.1 2×2 contingency table for family-based association studies based on a sample of N families with one affected child and both parents showing the matching of the two alleles of a parent. Consider a biallelic marker with alleles M_1, M_2. Small letters (a, b, c, d) denote allele counts. $2N$ denotes the total number of parental genotypes (i.e., of pairs of transmitted and non-transmitted alleles) to the affected child from the $2N$ parents

	Non-transmitted allele of one parent		
Transmitted allele of one parent	M_1	M_2	Total
M_1	a	b	$a+b$
M_2	c	d	$c+d$
Total	$a+c$	$b+d$	$2N$

(Spielman et al. 1993). The TDT is a haplotype-based analysis of the matched sample (Table 38.1). The test statistic is

$$\text{TDT} = (b-c)^2/(b+c).$$

The TDT compares whether the M_1 allele is more often transmitted to an affected child (b) than the M_2 allele (c) from heterozygous parents or vice versa. The test only considers M_1M_2 parents, since homozygous parents are not informative for preferential transmission of either allele.

The literature on family-based association analysis is vast (see, e.g., Whittaker and Morris (2001)). Important extensions of the above methods allow the application to multi-allelic markers, to tightly linked loci, and to quantitative traits. In addition, the design also allows for other types of nuclear families, such as sibships with affected and unaffected individuals (Spielman and Ewens 1998; Laird and Lange 2006). If a particular mode of inheritance is suspected, specialized versions of the TDT or related likelihood methods may yield higher power (Schaid 1999). If a candidate gene is to be investigated in detail, then a haplotype analysis can be carried out considering several biallelic polymorphisms (SNPs) in the same gene. The first step in a haplotype analysis is the estimation of the haplotype frequency in a population or the estimation of the most probable haplotype pair in an individual. For cohort or case-control studies, see Excoffier and Slatkin (1995), Stephens and Donnelly (2003), and Browning and Browning (2009); for family samples, see Rohde and Fuerst (2001) and Qian and Beckmann (2002). In the second step, linkage disequilibrium is investigated on the basis of the estimated haplotypes or haplotype frequencies. Some of these LD measures have already been described above (Devlin and Risch 1995).

Besides the analysis of main effects, gene-gene and gene-environment interactions can also be investigated in association analysis using standard tools. For gene-environment interaction, the case-only design enables the analysis of multiplicative interactions of factors, required to be independent in the population, on the basis of a sample of diseased individuals (Albert et al. 2001). If the independence assumption is valid, it is very efficient; if it is violated, such as for smoking-addiction genes and smoking, results can be severely distorted.

38.7 Current Methods and Outstanding Challenges

Genetics – and all its subdisciplines – has been an amazingly fast-moving field, with outstanding developments in both technology and biological insights. Methodological developments in statistical genetics and bioinformatics have had to scramble to keep pace; not the least of these challenges has been the daunting computational challenges posed by the massive data sets these advances have provided. In the remainder of this chapter, we review the current state of the art in the design of modern genetic epidemiology studies and the analysis of ultra-high-dimensional data and attempt to anticipate some of the novel developments that will be required in the foreseeable future.

38.7.1 Genome-Wide Association Studies

Fifteen years ago, Risch and Merikangas (1996) published a farsighted article in *Science* on the failure of traditional linkage analysis to uncover the genetics of complex diseases and made the then-radical suggestion that it would soon be possible to explore the entire genome by direct association methods. Their prediction came true, enabled by two developments in particular. The first was the advent of chip-based genotyping platforms that made it possible to assay hundreds of thousands to millions of SNP genotypes at a cost of under $1,000 per sample with high reliability. The second was a concerted effort by the public and private sectors to map the entire human genome (the Human Genome Project) and then to assemble a catalogue of known variants in a sample of Caucasian, African, and Asian populations (the International HapMap Project). In combination, these two advances provide a feasible way to directly genotype a large proportion of common variants and to predict many variants located next to the genotyped markers that are not typed directly.

It was nearly a decade before the first success of this approach was published in the form of a trio of papers on age-related macular degeneration in *Science*, one using this approach (Klein et al. 2005), which implicated a gene *CFH* in the complement pathway, a finding that has subsequently been confirmed numerous times. Since then, associations of about 210 diseases or quantitative traits and 1,300 genetic loci at genome-wide levels of significance ($p < 5 \times 10^{-8}$) have been published that have been independently replicated; the most recent version of the catalogue of published GWAS is available at the website of National Humane Genome Research Institute (NHGRI), National Institutes of Health (Hindorff et al. 2012; see also Hindorff et al. 2009). Study design and statistical analysis methods for such genome-wide association studies (GWAS) have been discussed in greater depth than is possible here (Hirschhorn and Daly 2005; Wang et al. 2005; Kraft and Cox 2008; McCarthy et al. 2008; Thomas 2010a, b; Witte 2010). Instead, we briefly review some of the recurring themes.

Multistage study designs. Early on, it was recognized that the multiple comparisons burden might be alleviated by some kind of staged design, in which only a portion of the sample would be used for screening the entire genome using one of the expensive high-density commercial platforms offering a fixed array of SNPs, followed by genotyping the remainder of the sample using a custom panel of only the most promising SNPs (Satagopan and Elston 2003). The final analysis then combined the information from both stages rather than treating it as a discovery and independent replication design (Skol et al. 2006); furthermore, it was possible to optimize the allocation of subjects to the two (or more) stages and the selection of the threshold for selecting SNPs to be genotyped in the later stages (Wang et al. 2006; Skol et al. 2007). With rapidly declining costs and increasing density of coverage of commercial panels, the interest in multistage designs has declined as investigators have recognized the advantage of having genome-wide data on the entire available sample as a resource for testing a broad range of hypotheses. However, the basic concept remains in spirit in the requirement for independent replication, as discussed in Sect. 38.7.3, and with the need for selecting a manageable number of individuals for next-generation sequencing technologies (Thomas et al. 2009).

Multiple testing and replication. With hundreds of thousands, if not millions, of associations being tested in a single study, there is an obvious need to avoid false-positive claims by adopting a stringent level of significance. A simple Bonferroni correction for one million SNPs would suggest a threshold of $0.05/10^6 = 5 \times 10^{-8}$, which has become the conventional criterion for claiming genome-wide significance, nominally ensuring a 5% probability of making at least one false-positive (family-wise error rate). This calculation fails to take into account the correlation among these tests due to linkage disequilibrium but is not a bad approximation for even the more recent platforms that allow testing of 2.5–5 million SNPs, as 1 million turns out to be roughly the effective number of independent tests in populations of European descent (or roughly double that in African-descent populations) (Pe'er et al. 2008). Fast asymptotic approximations have been developed (Conneely and Boehnke 2007) that allow for LD within a region, or permutation tests can be used as a gold standard for more complex dependency structures.

Despite the stringency of the genome-wide significance threshold, there are many factors than can lead to increased false-positives, some of which are discussed further below. But the key point is that a single study, however significant, is not usually considered convincing evidence of a genuine association without independent replication (Ioannidis 2007). Such replication is not simply to guard against chance variation (which can always be avoided simply by adopting an even more stringent significance threshold) but due to various sources of uncontrolled bias. Hence, real scientific replication should involve some elements of validation in different populations, by different investigators, using different methods. This is already demanded by the standard Bradford-Hill criteria in epidemiology used to establish validity of the association and strengthen belief in causality. This is not always possible in practice, however, such as for rare diseases where the discovery comes from a consortium of virtually all the available data or for studies conducted in unique settings (suggesting a need for some flexibility in the replication demands of granting agencies and journal editors!). Here in particular, the support by additional biological evidence is highly warranted.

Population substructure and study designs. One of the most pervasive sources of bias in GWAS is population substructure, as discussed in Sect. 38.6. Most GWAS therefore adjust instead for either an estimate of global ancestry from a finite set of founding populations using ancestry informative markers, typically using the STRUCTURE program (Pritchard et al. 2000) or using the top principal components from all or a subset of the markers, typically using the EIGENSTRAT program (Patterson et al. 2006). Either of these approaches tends to be quite effective at controlling the overall false-positive rate at least in homogeneous populations such as those of European descent. As a diagnostic for whether residual overdispersion due to uncontrolled population substructure remains after such adjustment, Quantile-Quantile plots of observed versus expected p-values of the single marker test statistics are generally used, and the genomic control overdispersion factor (Devlin and Roeder 1999) is checked to see if it is close to one. Control of population substructure can be more difficult in admixed populations like African-Americans and Hispanics, but these offer the advantage of being able to use within-individual comparisons for admixture mapping (Patterson et al. 2004; Freedman et al. 2006). For conventional GWAS scans in admixed populations, further adjustment for local ancestry (i.e., the ancestral origins of individual's chromosomes in the specific region of interest) may be necessary; the LAMP (Sankararaman et al. 2008) and HAPMIX (Price et al. 2009) programs can be used for this purpose.

Imputation. From the beginning, it has been understood that most associations discovered in a GWAS were unlikely to be directly causal, because only a small fraction of the genetic variation in the genome was being tested, but would hopefully reflect indirect associations with nearby causal variants through linkage disequilibrium between the causal and the marker loci. With the availability of the much more extensive catalogue of variation from the HapMap project, it has now become possible to infer the genotypes for most of the common variants in the genome by using imputation techniques from one of these standard reference panels (Li et al. 2010; Marchini and Howie 2010). Although programs such as MACH provide an assessment of the most likely genotype at each untyped locus, their use in association analysis would fail to account for the uncertainty in these imputations, essentially a form of measurement error leading to biased tests. A more appropriate procedure is therefore to use the estimated genotype probabilities or (under an additive model) the expected "gene dosage" as a continuous variable in a logistic regression analysis (Hu and Lin 2010).

Reporting. Given the potential range of problems and the approaches different investigators might take to addressing them, there is a need for some systematic guidance for how GWAS should be reported, if only to avoid subsequent problems in synthesizing the literature. Several authoritative statements have been issued by various groups to address this need, without imposing a straightjacket that would stifle investigators' creativity (Ehm et al. 2005; Chanock et al. 2007; Ioannidis et al. 2008; Hudson and Cooper 2009; Khoury et al. 2009; Little et al. 2009).

GWAS Summary. Over roughly the last 5 years, a consensus has emerged that most of the discovered novel associations of common, complex disorders with common SNPs have been relatively weak, with odds ratios typically in the range of about 1.2–1.6. Furthermore, even in the aggregate, these associations account for only a small portion of the total heritability estimated from classical twin or family studies (McCarthy and Hirschhorn 2008). While it is possible that these heritability estimates may be somewhat biased, it is certain that there remains a large portion of undiscovered genetic variation ("dark matter") to be accounted for (Hindorff et al. 2009; Manolio et al. 2009). Even for such a strongly related genetic and well-measured quantitative trait as height, the 180 loci that have so far been discovered based on meta-analysis of studies totaling over 183,000 individuals account for only about 10% of the total variability. Furthermore, it has been estimated that even with astronomical sample sizes, the number of loci of comparable effect sizes might rise to 600 but would still account for only about 20% of the heritability, which has been estimated at greater than 80% of the total phenotypic variation (Lango Allen et al. 2010). A variety of hypotheses have been advanced to account for this unexplained heritability, including rare variants, copy number variants, gene-environment and gene-gene interactions, and epigenetic effects, which will feature prominently as we move into the "post-GWAS" era.

38.7.2 Post-GWAS

Meta-analysis. Given the small effect sizes being sought and the enormous multiple comparisons penalty, most successful GWASs have required thousands of subjects. Nevertheless, as the experience with height demonstrates, no one study is likely to uncover more than a small fraction of the loci involved in a complex trait, and even larger sample sizes will be needed. Hence, the field has moved into a "Big Science" era, in which many investigators studying a given trait have formed consortia to pool all the available data for analyses of tens or hundreds of thousands of subjects (de Bakker et al. 2008; Zeggini and Ioannidis 2009). Some of these consortia have functioned simply to meet the replication requirement for each other's discoveries, but the more important purpose is to try to identify additional and weaker associations through a much larger sample size. This could in principle be accomplished by either a reanalysis of their combined raw data ("mega-analysis") or by meta-analysis of their summary statistics (Lin and Zeng 2010). In practice, the latter is usually much easier to accomplish, particularly if the different studies have used different genotyping platforms but can effectively impute genotypes for a larger, common set of HapMap SNPs (Zaitlen and Eskin 2010).

Fine mapping. Having identified one or more genome-wide significant and replicated regions, one might proceed to try to localize the region where a causal variant or variants might lie before attempting deep sequencing or functional studies. Here, the obvious strategy is simply to retest the available samples (or better, additional samples) with a higher density of markers in the region(s) of interest, but there are obvious trade-offs between sample size, number of regions that can be fine-mapped, region sizes, and density of markers (or criteria for selecting specific markers). These issues are amenable to methodological research, but there does not seem to be any generally agreed guidelines as of this writing. Given the speed the field is moving and the rapidly dropping costs of sequencing, many groups have decided to proceed directly to sequencing, bypassing the intermediate step of fine mapping.

Interactions and pathway analyses. Following an initial scan for main effects of SNPs (in either a single study or meta-analysis of several), much more remains to be explored. One possibility is that there could be larger gene-environment or gene-gene interaction effects that do not produce significant marginal effects. The obvious problem is that the number of possible interactions can be very much larger than the number of main effects: For a GWAS of one million SNPs, for example, there are half a trillion possible pairwise interactions. A simple Bonferroni correction for multiple comparisons would thus require a significance level of 1×10^{-13}, and of course, interaction tests would require much larger sample sizes than main effects even at the same significance level (Marchini et al. 2005). Power can be enhanced by "case-only" analyses based on an assumption of independence of the interacting factors in the source population. For example, a reanalysis of cleft palate data obtained substantially narrower confidence limits on the interaction between smoking and the *TGFα* gene (odds ratio $(OR) = 5.14$, 95% confidence interval $(\mathrm{CI}) = (1.68,15.7)$ for the case-only analysis compared with $OR = 6.57$, 95% $\mathrm{CI} = (1.72,20.0)$ for the case-control analysis), equivalent to a 30% reduction in sample size required for the same precision (Umbach and Weinberg 1997). Case-only analyses are, however, biased if this assumption is violated, as might arise due to LD among nearby pairs of loci, population stratification, or behavioral factors that induce an association between genes and environmental factors. To overcome these difficulty, various staged or hybrid approaches have been introduced (Evans et al. 2006; Kooperberg and Leblanc 2008; Mukherjee and Chatterjee 2008; Li and Conti 2009; Murcray et al. 2011). Although power will still be much lower than for main effects, these various methods generally yield much better power than a simple exhaustive search (Cornelis et al. 2012; Mukherjee et al. 2012). To make systematic study of gene-environment interactions of adequate sample sizes possible in the future, it is essential that investigators planning new GWAS design their studies to have appropriate environmental measurements and appropriate population-based sampling schemes. For example, the recent US National Institutes of Health (NIH) "post-GWAS" initiative aimed at synthesizing all the available data on five cancer sites, replicating findings, and characterizing genetic risks and their modification by environmental exposures has been limited by the fact that many of the available studies have not collected any environmental exposure data, and if collected, the choice of measurements was highly variable and ranged from very crude to very detailed assessment.

Another possibility is that there could be many SNPs that individually fail to be genome-wide significant but that jointly contribute to a common pathway. A variety of methods have been developed for identifying subsets of genes in known pathways that collectively are overrepresented among the top GWAS associations. Of these, the technique of Gene Set Enrichment Analysis (Wang et al. 2010), originally developed for gene expression data, has been most widely used. Hierarchical Bayes methods provide a more flexible regression-based approach to incorporate external knowledge about genomic, pathway, or functional annotation into the analysis of GWAS data (Lewinger et al. 2007). Cantor et al. (2010) provide an extensive review of these and other approaches to prioritizing GWAS associations for further investigation. Key to all these methods is the extent and quality of external databases such as the Kyoto Encyclopedia of Genes and Genomes and the Gene Ontology, which can be used for annotation in a systematic manner (Thomas et al. 2007). It is an interesting phenomenon that, after the failure of most candidate gene studies to yield replicable findings, the enthusiasm for the "agnostic" GWAS approach is now drifting back toward a synthesis of pathway-based and agnostic reasoning!

Functional studies. A somewhat unexpected finding from many GWAS is how few of the discovered associations lie in coding regions of genes (Hindorff et al. 2009). While some of the SNP associations could reflect LD with nearby coding variants that have not yet been discovered, it seems more likely that the majority will reflect variants in promoter regions of genes or long-range enhancers. Molecular techniques for functional characterization of causal associations will depend on the nature of the postulated effect, a topic which is beyond the scope of this article; for a recent set of recommendations, see Freedman et al. (2011). Nevertheless, there is a growing interest in "integrative genomics" approaches that can combine information across different types of data, such as SNPs, mutations in tumor tissues, transcriptomics, methylomics, metabolomics, and proteomics (Schadt et al. 2005; Hawkins et al. 2010).

38.7.3 Targeted, Whole-Exome, and Whole-Genome Sequencing

GWASs are based on an underlying "common disease, common variant" hypothesis (Reich and Lander 2001), which postulates that complex diseases are caused, at least in part, by common variants that can be effectively tagged by other SNPs in the region. The SNP panels used in most GWAS, based on one million or fewer SNPs, indeed are generally effective at tagging most "common" variants (conventionally defined as those with minor allele frequencies (MAF) of at least 5%). Newer generations of 2.5–5 million SNP panels will enhance the coverage of "uncommon" variants (those in the range of 1–5% MAF), but even these are not expected to provide good coverage of "rare" variants (less than 1% MAF). To discover these, direct sequencing will be necessary. The advent of several different massively parallel and fast "next-generation sequencing" (NGS) technologies (Davey et al. 2011) has now made this cost-effective, at least for targeted regions such as those around selected GWAS hits or the whole-exome. At the time of this writing, costs are typically around $1,000 per sample for whole exome sequencing and $5,000 for the whole genome, but the "$1,000 genome" (Mardis 2006) is anticipated in the near future. However, the data management and data analysis challenges are formidable. In addition to the storage problems of terabytes of raw data produced for even a single subject (petabytes for a typical study), NGS does not directly yield genotype calls but rather a random set of short sequence reads that must be aligned to cover the region of interest. The number of reads at any specific location is thus randomly distributed (with a mean given by the average depth of coverage), so the genotype can only be inferred probabilistically, taking into account possible errors in the reads themselves. Thus, in addition to the trade-offs mentioned earlier about region size and sample size, a further dimension to the design challenge is the necessary depth of sequencing, subject to a constraint on total cost (Sampson et al. 2011). The optimal design will depend upon the purpose – whether for discovery of rare variants or for testing association. One promising option is to perform the sequencing on pooled DNA samples (Futschik and Schlotterer 2010), which can considerably reduce costs, but this adds complexity to the optimization of the numbers of subjects and numbers of pools.

Multiple Rare Variant Analyses. One of the first analytical issues that arises concerns testing of association with rare variants. For feasible sample sizes, it is virtually impossible to test association with any single rare variant, both because of their rarity and the massive multiple comparisons penalty – for a typical whole-genome sequencing study, for example, one might expect to discover on the order of 20 million variants, most appearing only a few times. Interest has therefore tended to shift to tests of the "multiple rare variants" hypothesis (Price et al. 2010). The most commonly used technique is some form of "burden" test, which simply compares the total number of rare variants at a particular locus carried by cases and controls, possibly weighted in some fashion by their frequency or other characteristics (see Basu and Pan (2011) for a comparison of the available methods), but hierarchical Bayes methods that take account of model uncertainty offer a flexible and attractive alternative (Quintana et al. 2011).

Study designs for assessing causality. Most GWASs have used a case-control design with unrelated subjects for greatest statistical power. However, family-based designs offer two important advantages for studying rare variants. First, the sample can be enriched for rare variants by targeting cases with a strong family history. But more important is that looking at the pattern of cosegregation of variants with disease within families offers great potential for distinguishing causal variants from private polymorphisms that are simply circulating in the family but have nothing to do with the disease (Zhu et al. 2010; Shi and Rao 2011). This is essentially a form of linkage analysis, which has long been recognized as the design of choice for mapping rare major genes.

38.7.4 Risk Models and Translational Significance

Finally, as the number of genetic associations grows, it is natural to ask whether they can be used for genetic risk prediction. For genetic risk prediction, the purpose for the test must be distinguished, for example, a screening test for the general population or a high-risk population or a confirmation or exclusion test for a particular mutation for a particular family member of a disease-enriched family.

For major genes, segregation models as described above, using available genetic and non-genetic information and possibly indirect LD measures, can be used for prediction. For screening in populations, clearly a single SNP with relative risk less than 1.5 has very little value, but in the aggregate, one might consider a risk index based on all the known variants. For prostate cancer, for example, there are now more than 50 known GWAS associations, and a substantial portion of the population will carry several of these variants. Unfortunately, attempts to do this so far have not been particularly encouraging (Jostins and Barrett 2011; MacInnis et al. 2011; So et al. 2011; Newcombe et al. 2012). For prostate cancer in African-Americans, Haiman et al. (2011) reported a twofold gradient in predicted risk across quartiles of risk using 40 GWAS SNPs and 3.5-fold using only the 27 that were significantly associated among African-Americans, but the latter comparison may be subject to some overfitting. A clinically useful screening test should have both high sensitivity and high specificity (Kraft and Hunter 2009; Kraft et al. 2009; Janssens et al. 2011). Since both depend on where one draws the line between "normal" and "elevated" risk, a widely used measure of the overall performance of a screening test is the Area Under the Receiver Operating Curve (AUC) (Sanderson et al. 2005; Zou et al. 2007), obtained by plotting sensitivity against one minus specificity across the range of possible cut-points of the index (here, the predicted genetic risk score). For breast and prostate cancer, Machiela et al. (2011) found AUCs of 0.53 and 0.57, respectively, compared with 0.50 for an index that was no better than chance. Clinically useful risk indices would require an AUC of the order of 0.80 or better. For Crohn's disease, this appears attainable (Pharoah et al. 2002), but not at this point for most cancers (Chatterjee et al. 2011). Perhaps a more important question is what is the additional predictive value of genetic test results on top of established risk factors, including family history (Pencina et al. 2008; So et al. 2011). With the proliferation of direct-to-consumer genetic testing kits, some based on rather flimsy scientific evidence or with somewhat misleading advice about possible lifestyle changes to improve their risks, this question becomes of immediate translational significance (Hudson et al. 2007; Kaye 2008). See Levy et al. (2007) for a discussion of the scientific and ethical significance of the publication of the first complete human genome sequence. (The United States Genetic Information Non-discrimination Act (May 5, 2008) and a similar law in Germany (Gendiagnostikgesetz vom 31. Juli 2009 (BGBl. I S. 2529, 3672)) were enacted to protect individuals from discrimination based on genetic test results.) Although genetic risk prediction for the general population may still be some ways off, there is potentially greater utility in predicting genetic variation in response to treatments because these effects are likely to be much stronger due to the lack of time for evolution to eliminate deleterious variants, considering the recentness of most drug exposures (Altshuler et al. 2008).

38.8 Conclusions

The field of genetic epidemiology is in the midst of a fundamental paradigm shift. Originally based on methods for describing familial aggregation, testing for the existence of a genetic basis (segregation analysis), and localizing genetic causes (linkage analysis), the mainstay has become testing association with directly measured genotypes. A decade ago, this was feasible only for a modest number of variants in candidate genes, an approach that is now widely viewed as not having been particularly rewarding because of our lack of success in picking good candidates. With the advent of high-density genotyping platforms, agnostic scans for common variants across the entire genome have become popular and have led to many unexpected discoveries, albeit generally with rather small effect sizes that even in the aggregate account for only a modest proportion of the total estimated heritability of most complex diseases. These data are now being mined with sophisticated algorithms in the hope of identifying novel pathways across many of the suggestive, if not genome-wide significant, associations. Future directions are aimed at trying to identify the cause of the remaining unexplained heritability by targeted, whole-exome, or (soon) whole-genome sequencing – technologies that will pose formidable statistical and computational challenges – and by understanding the biological basis of the observed associations through regulatory, epigenetic, or other mechanisms.

Acknowledgements Dr. Bickeböller is supported by the German Research Foundation (DFG) grants (GRK 1644/1, KFO 241). Dr. Thomas is supported by NIH research grants ES019876, ES07048, MH084678, and HG005927.

References

Ainsworth HF, Unwin J, Jamison DL, Cordell HJ (2011) Investigation of maternal effects, maternal-fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. Genet Epidemiol 35:19–45

Albert PS, Ratnasinghe D, Tangrea J, Wacholder S (2001) Limitations of the case-only design for identifying gene-environment interactions. Am J Epidemiol 154:687–693

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322:881–888

Basu S, Pan W (2011) Comparison of statistical tests for disease association with rare variants. Genet Epidemiol 35:606–619

Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84:210–223

Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am J Hum Genet 86:6–22

Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, Brooks LD, Cardon LR, Daly M, Donnelly P, Fraumeni JF Jr, Freimer NB, Gerhard DS, Gunter C, Guttmacher AE, Guyer MS, Harris EL, Hoh J, Hoover R, Kong CA, Merikangas KR, Morton CC, Palmer LJ, Phimister EG, Rice JP, Roberts J, Rotimi C, Tucker MA, Vogan KJ, Wacholder S, Wijsman EM, Winn DM, Collins FS (2007) Replicating genotype-phenotype associations. Nature 447:655–660

Chatterjee N, Park J-H, Caporaso N, Gail MH (2011) Predicting the future of genetic risk prediction. Cancer Epidemiol Biomark Prev 20:3–8

Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42:393–399

Conneely KN, Boehnke M (2007) So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated tests. Am J Hum Genet 81:1158–1168

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261:921–923

Cornelis MC, Tchetgen Tchetgen EJ, Liang L, Qi L, Chatterjee N, Hu FB, Kraft P (2012) Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. Am J Epidemiol 175:191–202

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510

Day NE, Simons MJ (1976) Disease susceptibility genes – their identification by multiple case family studies. Tissue Antigens 8:109–119

de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet 17:R122–R128

Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–337

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Devlin B, Risch N, Roeder K (1996) Disequilibrium mapping: composite likelihood for pairwise disequilibrium. Genomics 36:1–16

Ehm MG, Nelson MR, Spurr NK (2005) Guidelines for conducting and reporting whole genome/large-scale association studies. Hum Mol Genet 14:2485–2488

Elston RC (1995) Twixt cup and lip: how intractable is the ascertainment problem? Am J Hum Genet 56:15–17

Elston RC (1998) Methods of linkage analysis – and the assumptions underlying them. Am J Hum Genet 63:931–934

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigrees. Hum Hered 21:523–542

Evans DG, Harris R (1992) Heterogeneity in genetic conditions. Q J Med 84:563–565

Evans DM, Marchini J, Morris AP, Cardon LR (2006) Two-stage two-locus models in genome-wide association. PLoS Genet 2:e157

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927

Falk CT, Rubinstein P (1987) Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227–233

Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, Waliszewska A, Penney K, Steen RG, Ardlie K, John EM, Oakley-Girvan I, Whittemore AS, Cooney KA, Ingles SA, Altshuler D, Henderson BE, Reich D (2006) Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proc Natl Acad Sci 103:14068–14073

Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, De Biasi M, Carlson C, Duggan D, James M, Liu P, Tichelaar JW, Vikis HG, You M, Mills IG (2011) Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet 43:513–518

Futschik A, Schlotterer C (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics 186:207–218

Haiman CA, Chen GK, Blot WJ, Strom SS, Berndt SI, Kittles RA, Rybicki BA, Isaacs WB, Ingles SA, Stanford JL, Diver WR, Witte JS, Chanock SJ, Kolb S, Signorello LB, Yamamura Y, Neslund-Dudas C, Thun MJ, Murphy A, Casey G, Sheng X, Wan P, Pooler LC, Monroe KR, Waters KM, Le Marchand L, Kolonel LN, Stram DO, Henderson BE (2011) Characterizing genetic risk at known prostate cancer susceptibility loci in African Americans. PLoS Genet 7:e1001387

Haldane JBS (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet 8:299–309

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. Science 250:1684–1689

Hardy GH (1908) Mendelian proportions in a mixed population. Science 28:49–50

Hawkins RD, Hon GC, Ren B (2010) Next-generation genomics: an integrative approach. Nat Rev Genet 11:476–486

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci 106:9362–9367

Hindorff LA, MacArthur J (European Bioinformatics Institute), Wise A, Junkins HA, Hall PN, Klemm AK, Manolio TA (2012) A catalog of published genome-wide association studies. Available at http://www.genome.gov/gwastudies. Accessed 8 July 2012

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common disease and complex traits. Nat Rev Genet 6:95–108

Holmans P (1993) Asymptotic properties of affected-sib-pair linkage analysis. Am J Hum Genet 52:362–374

Hu YJ, Lin DY (2010) Analysis of untyped SNPs: maximum likelihood and imputation methods. Genet Epidemiol 34:803–815

Hudson K, Javitt G, Burke W, Byers P (2007) ASHG statement* on direct-to-consumer genetic testing in the united states. Obstet Gynecol 110:1392–1395

Hudson TJ, Cooper DN (2009) STREGA: a 'How-To' guide for reporting genetic associations. Hum Genet 125:117–118

IGES (2012) International Society for Genetic Epidemiology. http://geneticepi.org/front. Accessed 8 July 2012

Ioannidis JP (2007) Non-replication and inconsistency in the genome-wide association setting. Hum Hered 64:203–213

Ioannidis JP, Boffetta P, Little J, O'Brien TR, Uitterlinden AG, Vineis P, Balding DJ, Chokkalingam A, Dolan SM, Flanders WD, Higgins JP, McCarthy MI, McDermott DH, Page GP, Rebbeck TR, Seminara D, Khoury MJ (2008) Assessment of cumulative evidence on genetic associations: interim guidelines. Int J Epidemiol 37:120–132

Janssens AC, Ioannidis JP, van Duijn CM, Little J, Khoury MJ (2011) Strengthening the reporting of genetic risk prediction studies: the grips statement. Genet Med 13:453–456

Jostins L, Barrett JC (2011) Genetic risk prediction in complex disease. Hum Mol Genet 20:R182–188

Kaye J (2008) The regulation of direct-to-consumer genetic tests. Hum Mol Genet 17:R180–183

Khoury M, Beaty T, Cohen B (1993) Fundamentals of genetic epidemiology. Oxford University Press, Oxford

Khoury MJ, Bertram L, Boffetta P, Butterworth AS, Chanock SJ, Dolan SM, Fortier I, Garcia-Closas M, Gwinn M, Higgins JP, Janssens AC, Ostell J, Owen RP, Pagon RA, Rebbeck TR, Rothman N, Bernstein JL, Burton PR, Campbell H, Chockalingam A, Furberg H, Little J, O'Brien TR, Seminara D, Vineis P, Winn DM, Yu W, Ioannidis JP (2009) Genome-wide association studies, field synopses, and the development of the knowledge base on genetic variation and human diseases. Am J Epidemiol 170:269–279

Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor h polymorphism in age-related macular degeneration. Science 308:385–389

Kooperberg C, Leblanc M (2008) Increasing the power of identifying gene x gene interactions in genome-wide association studies. Genet Epidemiol 32:255–263

Kraft P, Cox DG (2008) Study designs for genome-wide association studies. Adv Genet 60:465–504

Kraft P, Hunter DJ (2009) Genetic risk prediction – are we there yet? N Engl J Med 360:1701–1702

Kraft P, Wacholder S, Cornelis MC, Hu FB, Hayes RB, Thomas G, Hoover R, Hunter DJ, Chanock S (2009) Beyond odds ratios – communicating disease risk based on genetic profiles. Nat Rev Genet 10:264–269

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363

Laird NM, Lange C (2006) Family-based designs in the age of large-scale gene-association studies. Nat Genet 7:385–394

Lalouel JM, Rao DC, Morton NE, Elston RC (1983) A unified model for complex segregation analysis. Am J Hum Genet 35:816–826

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci 84:2363–2367

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2048

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segre AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, Qi L, Vernon Smith A, Magi R, Pastinen T, Liang L, Heid IM, Luan J, Thorleifsson G, Winkler TW, Goddard ME, Sin Lo K, Palmer C, Workalemahu T, Aulchenko YS, Johansson A, Zillikens MC, Feitosa MF, Esko T, Johnson T, Ketkar S, Kraft P, Mangino M, Prokopenko I, Absher D, Albrecht E, Ernst F, Glazer NL, Hayward C, Hottenga JJ, Jacobs KB, Knowles JW, Kutalik Z, Monda KL, Polasek O, Preuss M, Rayner NW, Robertson NR, Steinthorsdottir V, Tyrer JP, Voight BF, Wiklund F, Xu J, Zhao JH, Nyholt DR, Pellikka N, Perola M, Perry JR, Surakka I, Tammesoo ML, Altmaier EL, Amin N, Aspelund T, Bhangale T, Boucher G, Chasman DI, Chen C, Coin L, Cooper MN, Dixon AL, Gibson Q, Grundberg E, Hao K, Juhani Junttila M, Kaplan LM, Kettunen J, Konig IR, Kwan T, Lawrence RW, Levinson DF, Lorentzon M, McKnight B, Morris AP, Muller M, Suh Ngwa J, Purcell S, Rafelt S, Salem RM, Salvi E, Sanna S, Shi J, Sovio U, Thompson JR, Turchin MC, Vandenput L, Verlaan DJ, Vitart V, White CC, Ziegler A, Almgren P, Balmforth AJ, Campbell H, Citterio L, De Grandi A, Dominiczak A, Duan J, Elliott P, Elosua R, Eriksson JG, Freimer NB, Geus EJ, Glorioso N, Haiqing S, Hartikainen AL, Havulinna AS, Hicks AA, Hui J, Igl W, Illig T, Jula A, Kajantie E, Kilpelainen TO, Koiranen M, Kolcic I, Koskinen S, Kovacs P, Laitinen J, Liu J, Lokki ML, Marusic A, Maschio A, Meitinger T, Mulas A, Pare G, Parker AN, Peden JF, Petersmann A, Pichler I, Pietilainen KH, Pouta A, Ridderstrale M, Rotter JI, Sambrook JG, Sanders AR, Schmidt CO, Sinisalo J, Smit JH, Stringham HM, Bragi Walters G, Widen E, Wild SH, Willemsen G, Zagato L, Zgaga L, Zitting P, Alavere H, Farrall M, McArdle WL, Nelis M, Peters MJ, Ripatti S, van Meurs JB, Aben KK, Ardlie KG, Beckmann JS, Beilby JP, Bergman RN, Bergmann S, Collins FS, Cusi D, den Heijer M, Eiriksdottir G, Gejman PV, Hall AS, Hamsten A, Huikuri HV, Iribarren C, Kahonen M, Kaprio J, Kathiresan S, Kiemeney L, Kocher T, Launer LJ, Lehtimaki T, Melander O, Mosley TH Jr, Musk AW, Nieminen MS, O'Donnell CJ, Ohlsson C, Oostra B, Palmer LJ, Raitakari O, Ridker PM, Rioux JD, Rissanen A, Rivolta C, Schunkert H, Shuldiner AR, Siscovick DS, Stumvoll M, Tonjes A, Tuomilehto J, van Ommen GJ, Viikari J, Heath AC, Martin NG, Montgomery GW, Province MA, Kayser M, Arnold AM, Atwood LD, Boerwinkle E, Chanock SJ, Deloukas P, Gieger C, Gronberg H, Hall P, Hattersley AT, Hengstenberg C, Hoffman W, Lathrop GM, Salomaa V, Schreiber S, Uda M, Waterworth D, Wright AF, Assimes TL, Barroso I, Hofman A, Mohlke KL, Boomsma DI, Caulfield MJ, Cupples LA, Erdmann J, Fox CS, Gudnason V, Gyllensten U, Harris TB, Hayes RB, Jarvelin MR, Mooser V, Munroe PB, Ouwehand WH, Penninx BW, Pramstaller PP, Quertermous T, Rudan I, Samani NJ, Spector TD, Volzke H, Watkins H, Wilson JF, Groop LC, Haritunians T, Hu FB, Kaplan RC, Metspalu A, North KE, Schlessinger D, Wareham NJ, Hunter DJ, O'Connell JR, Strachan DP, Wichmann HE, Borecki IB, van Duijn CM, Schadt EE, Thorsteinsdottir U, Peltonen L, Uitterlinden AG, Visscher PM, Chatterjee N, Loos RJ, Boehnke M, McCarthy MI, Ingelsson E, Lindgren CM, Abecasis GR, Stefansson K, Frayling TM, Hirschhorn JN (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467:832–838

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC (2007) The diploid genome sequence of an individual human. PLoS Biol 5:e254

Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC (2007) Hierarchical bayes prioritization of marker associations from a genome-wide association scan for further investigation. Genet Epidemiol 31:871–882

Li D, Conti DV (2009) Detecting gene-environment interactions using a combined case-only and case-control approach. Am J Epidemiol 169:497–504

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34:816–834

Lin DY, Zeng D (2010) Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genet Epidemiol 34:60–66

Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, Khoury MJ, Cohen B, Davey-Smith G, Grimshaw J, Scheet P, Gwinn M, Williamson RE, Zou GY, Hutchings K, Johnson CY, Tait V, Wiens M, Golding J, van Duijn C, McLaughlin J, Paterson A, Wells G, Fortier I, Freedman M, Zecevic M, King R, Infante-Rivard C, Stewart A, Birkett N (2009) Strengthening the reporting of genetic association studies (STREGA): an extension of the STROBE statement. PLoS Med 6:e22

Machiela MJ, Chen C-Y, Chen C, Chanock SJ, Hunter DJ, Kraft P (2011) Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. Genet Epidemiol 35(6):506–514

MacInnis RJ, Antoniou AC, Eeles RA, Severi G, Al Olama AA, McGuffog L, Kote-Jarai Z, Guy M, O’Brien LT, Hall AL, Wilkinson RA, Sawyer E, Ardern-Jones AT, Dearnaley DP, Horwich A, Khoo VS, Parker CC, Huddart RA, Van As N, McCredie MR, English DR, Giles GG, Hopper JL, Easton DF (2011) A risk prediction algorithm based on family history and common genetic variants: application to prostate cancer with potential clinical impact. Genet Epidemiol 35:549–556

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11:499–511

Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37:413–417

Mardis ER (2006) Anticipating the 1,000 dollar genome. Genome Biol 7:112

Maynard Smith J (1989) Evolutionary genetics. Oxford University Press, Oxford

McCarthy MI, Hirschhorn JN (2008) Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet 17:R156–165

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356–369

McKusick VA (1998) Mendelian inheritance in man. Catalogs of human genes and genetic disorders, 12th edn. Johns Hopkins University Press, Baltimore

Mendel GJ (1865) Versuche über Pflanzenhybriden. Verhandlungen des Naturforschenden Vereins, Brünn

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318

Morton NE, MacLean CJ (1974) Analysis of family resemblance. 3. Complex segregation of quantitative traits. Am J Hum Genet 26:489–503

Mukherjee B, Chatterjee N (2008) Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics 64:685–694

Mukherjee B, Ahn J, Gruber SB, Chatterjee N (2012) Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. Am J Epidemiol 175:177–190

Murcray CE, Lewinger JP, Conti DV, Thomas DC, Gauderman WJ (2011) Sample size requirements to detect gene-environment interactions in genome-wide association studies. Genet Epidemiol 35:201–210

Newcombe PJ, Reck BH, Sun J, Platek GT, Verzilli C, Kader AK, Kim S-T, Hsu F-C, Zhang Z, Zheng SL, Mooser VE, Condreay LD, Spraggs CF, Whittaker JC, Rittmaster RS, Xu J (2012)

A comparison of Bayesian and frequentist approaches to incorporating external information for the prediction of prostate cancer risk. Genet Epidemiol 36:71–83

Newman B, Austin MA, Lee M, King MC (1988) Inheritance of human breast cancer: evidence for autosomal dominant transmission in high-risk families. Proc Natl Acad Sci 85:3044–3048

OMIM (2012) Online Inheritance In Man. http://www.ncbi.nlm.nih.gov/omim/. Accessed 8 July 2012

Ott J (1999) Analysis of human genetic linkage, 3rd edn. Johns Hopkins, Baltimore

Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, Daly MJ, Reich D (2004) Methods for high-density admixture mapping of disease genes. Am J Hum Genet 74:979–1000

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2:e190

Pe'er I, Yelensky R, Altshuler D, Daly MJ (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol 32:381–385

Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 27:157–172; discussion 207–212

Penrose LS (1953) The general purpose sibpair linkage test. Ann Eugen 18:120–124

Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA (2002) Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet 31:33–36

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet 5:e1000519

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86:832–838

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67:170–181

Qian D, Beckmann L (2002) Minimum recombinant haplotyping in pedigrees. Am J Hum Genet 70:1434–1445

Quintana MA, Bernstein JL, Thomas DC, Conti DV (2011) Incorporating model uncertainty in detecting rare variants: the Bayesian risk index. Genet Epidemiol 35:638–649

Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17:502–510

Risch N (1984) Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes. Am J Hum Genet 36:363–386

Risch N (1991) A note on multiple testing procedures in linkage analysis. Am J Hum Genet 48:1058–1064

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1616–1617

Rohde K, Fuerst R (2001) Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information. Hum Mutat 17:289–295

Sampson J, Jacobs K, Yeager M, Chanock S, Chatterjee N (2011) Efficient study design for next generation sequencing. Genet Epidemiol. doi:10.1002/gepi.20575 (Epub ahead of print)

Sanderson S, Zimmern R, Kroese M, Higgins J, Patch C, Emery J (2005) How can the evaluation of genetic tests be enhanced? Lessons learned from the ACCE framework and evaluating genetic tests in the United Kingdom. Genet Med 7:495–500

Sankararaman S, Sridhar S, Kimmel G, Halperin E (2008) Estimating local ancestry in admixed populations. Am J Hum Genet 82:290–303

Satagopan JM, Elston RC (2003) Optimal two-stage genotyping in population-based association studies. Genet Epidemiol 25:149–157

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37:710–717

Schaid DJ (1999) Likelihoods and TDT for the case-parents design. Genet Epidemiol 16:250–260

Schaid DJ, Sommer SS (1994) Comparison of statistics for candidate-gene association studies. Am J Hum Genet 55:402–409

Sham P (1998) Statistics in human genetics. Wiley, New York

Shi G, Rao DC (2011) Optimum designs for next-generation sequencing to discover rare variants for common complex disease. Genet Epidemiol 35:572–579

Shute NC, Ewens WJ (1988) A resolution of the ascertainment sampling problem. II. Generalizations and numerical results. Am J Hum Genet 43:374–386

Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209–213

Skol AD, Scott LJ, Abecasis GR, Boehnke M (2007) Optimal designs for two-stage genome-wide association studies. Genet Epidemiol 31:776–788

So H-C, Kwan JSH, Cherny SS, Sham PC (2011) Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. Am J Hum Genet 88:548–565

Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–458

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169

Suarez BK, Hampe CL (1994) Linkage and association. Am J Hum Genet 54:554–559; author reply 60–63

Terwilliger JD, Ott J (1992) A haplotype based haplotype relative risk approach to detecting allelic associations. Hum Hered 42:337–346

Terwilliger JD, Ott J (1994) Handbook of human genetic linkage. Johns Hopkins University Press, Baltimore

Thomas D (2010a) Gene-environment-wide association studies: emerging approaches. Nat Rev Genet 11:259–272

Thomas DC (2010b) Design and analysis issues in genome-wide association studies. In: Khoury MJ, Bedrosian S, Gwinn M, Khoury MJ, Bedrosian S, Gwinn M, Higgins JPT, Ioannidis JPA, Little J (eds) Human genome epidemiology: building the evidence for using genetic information to improve health and prevent disease, 2nd edn. Oxford University Press, Oxford

Thomas DC, Casey G, Conti DV, Haile RW, Lewinger JP, Stram DO (2009) Methodological issues in multistage genome-wide association studies. Stat Sci 24:414–429

Thomas PD, Mi H, Lewis S (2007) Ontology annotation: mapping genomic regions to biological function. Curr Opin Chem Biol 11:4–11

Umbach DM, Weinberg CR (1997) Designing and analysing case-control studies to exploit independence of genotype and exposure. Stat Med 16:1731–1743

Vogel W (2000) Genetische Epidemiologie oder zur Spezifität von Subdisziplinen der Humangenetik. Med Genet 4:395–399

Wang H, Thomas DC, Pe'er I, Stram DO (2006) Optimal two-stage genotyping designs for genome-wide association scans. Genet Epidemiol 30:356–368

Wang K, Li M, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. Nat Rev Genet 11:843–854

Wang WYS, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6:109–118

Weinberg W (1980) Über den Nachweis der Vererbung beim Menschen. Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg 64:368–383

Weiss KM (1993) Genetic variation and human disease: principles and evolutionary approaches. Cambridge University Press, Cambridge

Whittaker JC, Morris AP (2001) Family-based tests of association and/or linkage. Ann Hum Genet 65:407–419

Whittemore AS, Tu IP (1998) Simple, robust linkage tests for affected sibs. Am J Hum Genet 62:1228–1242

Witte JS (2010) Genome-wide association studies and beyond. Annu Rev Publ Health 31:9–20 4 p following 20

Zaitlen N, Eskin E (2010) Imputation aware meta-analysis of genome-wide association studies. Genet Epidemiol 34(6):537–542

Zeggini E, Ioannidis JP (2009) Meta-analysis in genome-wide association studies. Pharmacogenomics 10:191–201

Zhou JY, Ding J, Fung WK, Lin S (2010) Detection of parent-of-origin effects using general pedigree data. Genet Epidemiol 34:151–158

Zhu X, Feng T, Li Y, Lu Q, Elston RC (2010) Detecting rare variants for complex traits using family and unrelated data. Genet Epidemiol 34:171–187

Zou KH, O'Malley AJ, Mauri L (2007) Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation 115:654–657

Directed Acyclic Graphs 39

Ronja Foraita, Jacob Spallek, and Hajo Zeeb

Contents

R. Foraita (✉)
Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

J. Spallek
Department of Epidemiology & International Public Health, University of Bielefeld – School of Public Health, Bielefeld, Germany

H. Zeeb
Department of Prevention and Evaluation, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

Faculty of Human and Health Sciences, University of Bremen, Bremen, Germany

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_65,

39.1 Introduction

A directed acyclic graph (DAG) can be thought of as a kind of flowchart that visualizes a whole causal etiological network, linking causes and effects. In epidemiology, the terms causal graph, causal diagram, and DAG are used as synonyms (Greenland et al. 1999). DAGs are considered to be of use for embedding causality in a formal causal framework (Hernán and Robins 2006; Robins 2001; Hernán et al. 2004). In probability theory, there is a somewhat different understanding of DAGs, which we will discuss later. This chapter aims to demonstrate how DAGs can help to formalize the search for answers to different research questions in epidemiology.

In the first section of this chapter, we review the application of DAGs in epidemiology. Data collected in the framework of observational epidemiological studies are usually insufficient in one way or another, for example, due to missing information, misclassification, or due to some other form of bias (see chapter ▸Basic Concepts of this handbook). Even when there is little doubt that an association found in observational epidemiological studies represents a true causal mechanism, such as in the case of smoking and lung cancer, information obtained may be biased, leading to a biased estimate of the strength of the association under study. A DAG is useful to determine confounding factors that bias the association between a risk factor and disease.

The second section of this chapter deals with probabilistic DAGs in the framework of Bayesian networks. Bayesian networks are graphical representations of a probability distribution over a set of variables. They are useful to identify the association structures among all variables under investigation and to visualize conditional relationships among these variables. Bayesian networks are appropriate in situations where we deal with complex or unknown association structures. Their key concept is to identify conditional independencies, where the study variables are symbolized as vertices in a DAG and edges between them reflect associations. That means, if two vertices are unconnected in the DAG, the corresponding variables in the probability model are conditionally independent given the remaining variables. We will introduce the probability theory behind Bayesian networks as well as data-driven inference to estimate the unknown parameters of the underlying statistical model. We also deal with model selection to search for the DAGs that best reflect the pattern of the data at hand. In the final sections, we discuss advantages and limitations of the DAG concept and give an outlook.

39.2 Causal Graphs: Concepts

Causal graphs associate events with other events. Causes necessarily precede their effects, and this temporal relation is part of the definition of a cause. Causal graphs provide a supportive approach to visualize causal links between exposures and outcomes in epidemiological research. Formal rules are used to develop the graphs and to derive appropriate analytical approaches. The theoretical framework for using DAGs in epidemiology has been developed only recently, motivating health scientists to work with DAGs in planning their investigations and analyzing their data. DAGs and probability theory can be incorporated into Bayesian networks to be applied to real data problems. Pearl (2009) specified the threefold role of graphs in probabilistic and statistical modeling as follows:

1. To provide convenient means of expressing assumptions
2. To facilitate economical representation of joint probability functions
3. To facilitate efficient inferences from observations

Causal graphs move away from purely mathematical language to an alternative way of communicating, understanding, and evaluating causality in empirical sciences.

As pointed out by Greenland et al. (1999), the theory of DAGs offers the benefit of compact graphical as well as probabilistic representations of assumptions, for example, about confounding, in epidemiological studies. In addition, the evaluation of confounding especially in the presence of multiple potential confounders is aided through a graphical approach when traditional textbook criteria of confounding fall short. Graph theory can help to identify the set of confounders for which adjustment is required, whereas other potential confounders can be discarded according to the DAG. Causal graphs should be understood as tools rather than as independent novel concepts. They help to clarify network structures and enforce thinking about the relationships between variables which may otherwise remain vague or totally omitted. For non-mathematicians, causal graphs provide an opportunity to derive conclusions from a diagram developed through the application of a set of formal rules (Glymour 2006).

39.2.1 History and Development

Causal graphs were first used almost a 100 years ago in the field of genetics and later in the social sciences, econometrics, physics, and many other areas. Graphs usually represent deterministic functional relationships between cause and effect, where probabilistic elements indicate unobserved variables in the equation. Linking a probabilistic perspective with the aim to deduce deterministic causal relationships from empirical data was at the core of developments in the field of causal graphs towards the end of the last century. A formal framework of the causal graph concept was first developed in the early 1990s, later published in a book by Spirtes et al. (2001). The underlying mathematical theory is that of Bayesian networks (see Sect. 39.3), which represent probabilistic assumptions about the associations between different components of the network. The rapid emergence of artificial intelligence systems since the 1980s was a major input to the development of Bayesian network theory. These systems allow for the efficient ordering and interpretation of new observations, searching for associations while accounting, both, for prior information and the new set of observations in a coherent way (Pearl 2009). This is not possible without advanced computer technology, as the amount of data, for example, in medical research, climate sciences, but also in the framework of epidemiological studies, grows rapidly.

39.2.2 DAGs and Other Methods of Causal Modeling

A prominent causal model in epidemiology is the sufficient-component cause model by Kenneth Rothman (see chapter ▶Basic Concepts of this handbook; Rothman 1976; Rothman et al. 2008). This model uses pie charts to represent causes, with the different pieces of the pie standing for component parts of the overall cause. Once the causal mechanism is deemed to be complete in the sense that the minimal set of conditions and events that lead to the effect are present, a cause is described as sufficient. All the components of the cause are required for the effect to occur. If a specific component needs to be always present in the set of component causes because the effect never occurs without it, this component is called necessary. The component approach is easily reconcilable with the common understanding of multifactorial diseases in modern medicine. While conceptually helpful to understand causation, the specification of sufficient-component cause models in defined situations requires in-depth knowledge of biological mechanisms or other information that are often not available (Greenland and Brumback 2002). Causal graphs as well as the sufficient-component cause model have their strength in a qualitative description of causal relations. Two other causal modeling approaches also build on aspects of qualitative description of causal relations but provide quantification of specific exposure-disease relations. These are the counterfactual (or potential outcomes) models (see Rothman et al. (2008) for more details) and structural equation models (Greenland and Brumback 2002).

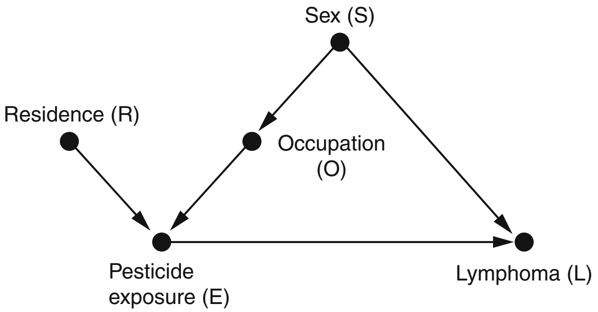
Within the realm of counterfactual models (Rubin 1974), g-estimation (Robins et al. 1992) and marginal structural modeling (Robins et al. 2000) are more recent developments. Structural equation modeling has a long tradition in the social sciences (Pearl 2009). One logical link between the different models is given through their focus on alternatives or contrasts in the representation, either graphical or probabilistic, of causal structures between exposures (or interventions) and outcomes.

39.2.3 DAGs in Practice: Some Introductory Notations

Directed acyclic graphs are directed in that they symbolize the associations between different components of a supposed causal network through arrows pointing from a cause to its effect. They are acyclic since no path of directed arrows is allowed to form a closed loop. There are other versions of graphs that employ undirected graphs and cyclical associations, but these are not of interest here since the investigation of causality implies direction from cause to effect (see Greenland et al. 1999). The principles of DAGs are illustrated by the following example which we will revisit throughout this chapter.

Example 39.1a.
Suppose we want to investigate the association between pesticides and the occurrence of lymphoma in adults employed in various industrial plants, with the underlying hypothesis that pesticide exposure increases the risk of lymphoma. In a graph, a line with an arrowhead

Fig. 39.1 DAG of the lymphoma example



pointing from *pesticide* (exposure) to *lymphoma* (outcome) symbolizes this situation. Further, we assume that *sex* affects exposure probability only through *occupation*, while *sex* is directly related to *lymphoma* risk as the outcome. *Residence* also affects *pesticide exposure* probability. Figure 39.1 gives the full picture for this simple situation and is used to introduce first definitions and concepts, which will be referred to later.

In explaining the notations used for constructing the graph, we refer to the basic paper by Greenland et al. (1999). A line connecting two variables in a causal diagram is called an arc or edge. Adjacent variables are those in direct neighborhood to each other, connected by an edge. A (causal) directed association is indicated by a single-headed arrow, pointing from cause to effect.

Example 39.1b.
In Fig. 39.1, the line linking *pesticide exposure* (E) and *lymphoma* (L) corresponds to an arrow, indicating in this example the assumption that pesticide exposure causes lymphoma, which in epidemiology translates to "modifies the risk of lymphoma," given the absence of bias and confounding.

Arrows, arcs, or edges end up in nodes or vertices, which are the points in the graph representing the variables considered, as, for example, *residence* (R), *pesticide exposure* (E), *occupation* (O), *sex* (S), and *lymphoma* (L) in our example. A directed path is a line made up of arrows all pointing in one direction and connecting adjacent vertices, for example, the path from R to L passing through E. When a path leads through a vertex, this vertex intercepts the path. A succession of edges that simply connect adjacent vertices, regardless of their direction, is simply called path – it is not directed. Ancestors or causes E and S of the variable L are those vertices on a directed path that leads into L. Conversely, L is named descendant of E and S. Looking only at two adjacent vertices, E is called a parent of L (child) if an arrow originates in E and leads to L.

A more general notation for DAGs denotes X as exposure, Y as outcome, and $Z_1, \ldots, Z_K$ as measured covariates. This will be used in Sect. 39.2.6 where the essential steps for the construction and interpretation of graphs are outlined.

39.2.4 DAGs and Confounding

An exposure-outcome association can be confounded by a third factor if this factor independently causes the outcome and is associated with the exposure under study. Failure to adjust for such confounding factors may lead to bias (see chapter ▸Confounding and Interaction of this handbook). Factors that are intermediate steps on the assumed causal path between exposure and outcome should, however, not be adjusted for. There are also situations, as initially pointed out by Weinberg (1993), where a factor that changes the exposure-outcome association and thus appears to be a confounder should not be adjusted for as this would actually introduce bias in the risk measures. This may be the case if this third factor is itself partly caused by the exposure. Furthermore, in specific situations, adjustment for a confounding factor may in turn introduce confounding by another factor previously not regarded as confounder of the exposure-outcome association under study.

DAGs help to graphically describe such situations and call for explicit consideration of the factors to be included (or not) in analyses of epidemiological studies. DAGs can therefore be applied as tools to identify variables that have to be adjusted for in order to attain an – ideally – unbiased estimate of the association between an exposure and an outcome.

39.2.5 Constructing a DAG to Identify and Control for Confounding

Before starting to build a DAG, one needs to assess which variables might be important for answering the research question. The choice of potentially relevant variables, for example, to explain the etiology of a disease, is based on one or several of the following considerations:

- Subjective assumptions, common sense
- Other studies, previous research
- Other information (e.g., theoretical considerations, expert opinions)

A DAG provides a formal way to organize the covariates based on the best possible assumption about the causal structure of these variables. The DAG leads to one (or several) so-called "minimally sufficient adjustment set" of covariates to be adjusted for in order to obtain an unconfounded estimation of the association. The following introduction to the rules for constructing a DAG is based on Glymour and Greenland (2008), Greenland et al. (1999), Hernán et al. (2002), and Glymour (2006).

39.2.6 Formal Rules for Constructing a DAG

As mentioned before, each arrow in a directed acyclic graph has only one arrow head and each arrow points from one variable to another variable, without closed loops. If there is an unmeasured (latent) variable that might influence the causal association between exposure and outcome, it has to be represented in the DAG as an unmeasured covariate. In most epidemiological studies, weak associations are unlikely to introduce major bias. Thus, the influence of these variables on the causal pathway is probably negligible, and therefore, omission of these variables from the DAG is recommended. The following notations are used to construct a DAG in order to identify confounders that have to be controlled for:

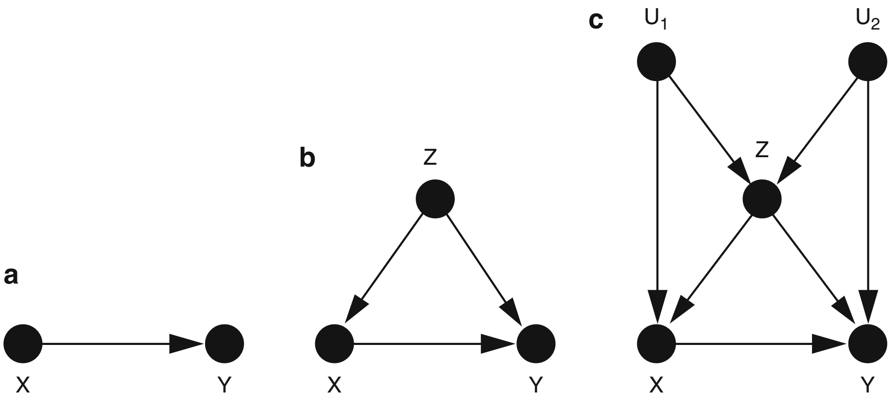


Fig. 39.2 (**a**) Exposure-outcome relationship (Step 1); (**b**) add measured covariates (Step 2); (**c**) add unmeasured covariates. X: *first delivery at older age* (exposure); Y: *breast cancer* (outcome); Z: *family history* (measured covariate); U_1: *SES* (unmeasured covariate); U_2: *BRCA I/II* (unmeasured covariate) (Step 3)

- Exposure: X
- Outcome: Y
- Measured covariates: $Z_1, Z_2, \ldots, Z_K$
- Unmeasured covariates: $U_1, U_2, \ldots, U_L$

Every covariate is a variable and not the manifestation of a variable, e.g., sex (male/female) and not male. For every association, a directed arrow is drawn.

Example 39.2.
We will explain the construction of a DAG step by step using a hypothetical study of the association between the dichotomous variable *first delivery at older age* (X) and risk for *breast cancer* (Y).

1. **Draw exposure and outcome:** In the first step the exposure-outcome relationship is constructed (see Fig. 39.2a). Remember that,
   - Each vertex is a variable
   - The graph must be directed
   - The graph must be acyclic.

   In Fig. 39.2a the exposure X is *first delivery at older age* pointing to the outcome Y, which is *breast cancer.*
2. **Include measured covariates** *Family history* of breast cancer (Z) is a measured covariate which might influence (i) *first delivery at older age* (i.e., women with a family history of breast cancer might become pregnant at a younger age) and (ii) the risk of *breast cancer* (see Fig. 39.2b).
3. **Include unmeasured covariates** In our example, *socioeconomic status* (SES) and *breast cancer early onset gene type I or II* (*BRCA I/II*) are unmeasured covariates (U_1, U_2, respectively). *SES* is associated with *first delivery at older age* and might be associated with *family history* due to a higher proportion of women who were screened for a family history in higher SES groups or due to another reproductive behavior in higher SES groups, while *BRCA I/II* is associated with *family history* and *breast cancer* (see Fig. 39.2c).
4. **Prepare a list of all covariates affected by the exposure** In this step all children of X are listed. Children are those vertices where the arrow from X is heading to. In this

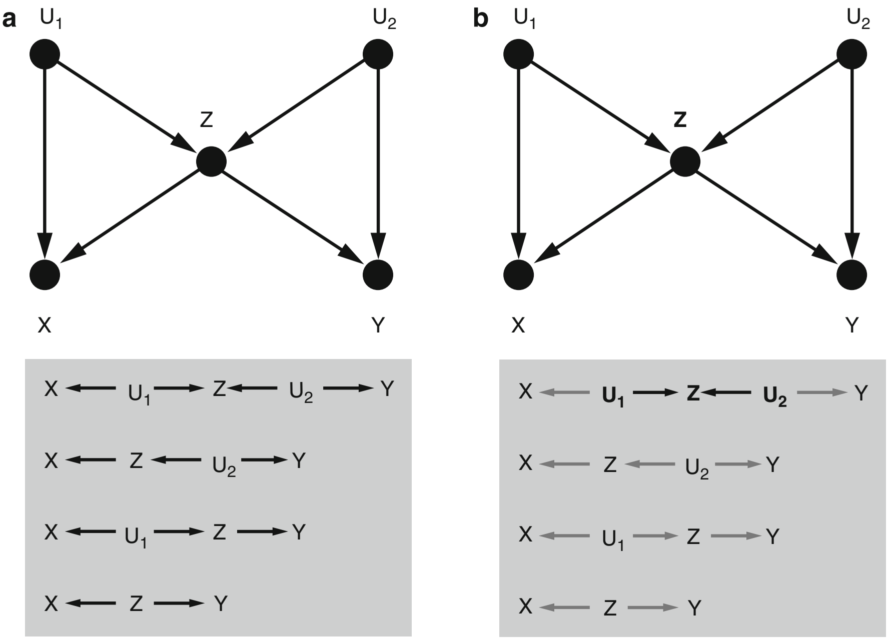


Fig. 39.3 (**a**) A list of backdoor paths after eliminating the exposure-outcome relation (Steps 4 and 5); (**b**) finding colliders (Step 6)

example, there is only one arrow originating in X: $X \rightarrow Y$. Eliminate this arrow to remove (for now) all assumed direct exposure effects (see Fig. 39.3a).

5. **Find all backdoor paths** For this purpose, arrow directions are not considered. Confounding is present if exposure and outcome are still connected after the elimination of all direct exposure effects. In this example, four pathways from X to Y exist after elimination of the direct exposure effects. These open pathways from X to Y are called backdoor paths (see Fig. 39.3a).
6. **Disregard the backdoor paths already blocked by a collider** Colliders are "collisions" of two paths, identifiable by covariates that have two arrows pointing towards them. Z is a collider in the example, because U_1 and U_2 point to Z (see Fig. 39.3b). Backdoor paths containing a collider are blocked because the pathway ends in the collider (here Z) and not in the outcome Y, and thus this backdoor path can be erased. In our example the first backdoor path $X \leftarrow U_1 \rightarrow Z \leftarrow U_2 \rightarrow Y$ can be deleted.
7. **Identify sufficient adjustment sets** List all unblocked, open backdoor paths and find a set of covariates that would block all still existing open backdoor paths by adjusting for this set. In our example the minimal sufficient adjustment set (MSAS) would include only Z. Adjusting for Z would block all open backdoor paths (see Fig. 39.4).
8. **Looking for associations induced by adjustment (or stratification)** Adjusting for the covariates that were identified as a MSAS might introduce new associations between covariates that have not been present before adjustment. In our example, no association between U_1 (*SES*) and U_2 (*BRCA I/II*) was existing prior to adjustment. However, adjusting for Z (*family history*) introduces an association between these two covariates, for example, a higher risk for carrying *BRCA I/II* among persons with a high *SES*. This phenomenon of an induced spurious association was first described by Berkson (1946).

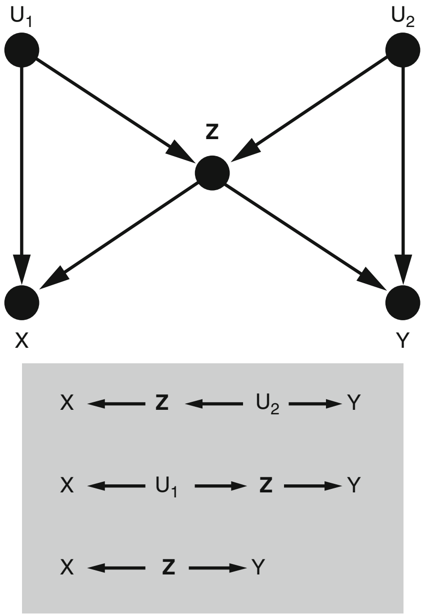


Fig. 39.4 Find potentially sufficient adjustment sets (Step 7)

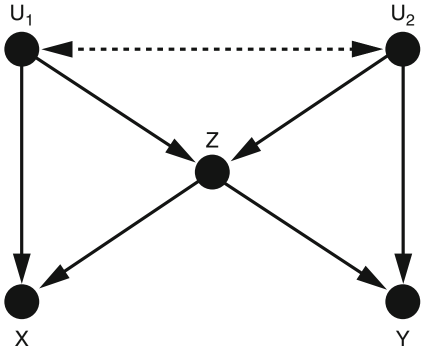


Fig. 39.5 Induced associations are expressed as a *dotted line* with *doubly headed arrow*, here between U_1 and U_2. The two minimal sufficient adjustment sets are $\{U_1, Z\}$ and $\{U_2, Z\}$ (Step 8)

The newly induced association is drawn as a dashed arrow with two heads while introduces a new, unblocked backdoor path (see Fig. 39.5).

9. **Repeat Steps 5–7 and find a new minimal sufficient adjustment set, preferably composed of measured covariates.** In our example, the final minimal sufficient adjustment sets are $\{U_1, Z\}$ or $\{U_2, Z\}$. Adjusting for U_1 and Z or U_2 and Z results in a situation in which the association between X and Y is not confounded by Z, U_1 or U_2.

The DAG does not allow to assess the quality or strength of the influence of a confounder; it only allows to state whether a covariate is a confounder or not. Drawing a DAG might, as in our example, end with the uncomfortable situation that in the minimal sufficient adjustment set, an unmeasured covariate is included. Inclusion of this covariate in statistical analyses is not possible. Generally, MSAS that include unmeasured covariates do not allow to correctly estimate the association between exposure and outcome.

DAGs help to determine whether a set of measured variables is sufficient for analyzing the association under study, for example, for identifying the causal effect of first delivery at older age on breast cancer risk. When there are confounding paths between exposure and outcome, it may be possible to close these paths by adjusting for other variables.

In practice, there may be several distinct sufficient sets and even several distinct minimal sufficient sets for controlling confounding. Researchers may sometimes wish to adjust for more variables than necessary. Identifying an MSAS can be valuable, because adjusting for more variables than needed implies the risk of introducing confounding and reducing precision. In addition, measuring unnecessary variables can be difficult and expensive.

Although the MSAS can quickly be identified by paper and pencil for small graphs, it can be a laborious task for complex graphs. Computer programs may help to identify the appropriate adjustment sets. `DAGitty` (Textor et al. 2011; Textor 2012) is an easy-to-use browser-based program; the `DAG` program (Knüppel and Stang 2010; Knüppel 2011) works with input and output files but cannot visualize the DAG; the `R` package `dagR` (Breitling 2010) provides a set of `R` functions to find the MSAS, however, also without graphical visualization.

39.3 Probabilistic DAGs and Bayesian Networks

In the previous sections, we introduced DAGs as a visual tool to identify a minimal sufficient adjustment set of factors that might bias the association in order to assess causal relationships. In this section, we consider probabilistic DAGs that describe probability statements and associations between random variables but without claiming to infer causality.

The key concept of probabilistic graphs is to interpret graphs in terms of conditional independencies and probabilities. In the following, we will focus on Bayesian networks as probabilistic DAGs in order to perform estimation and model selection. Bayesian networks are probability models to analyze and visualize the usually complex association structure that underlies any multivariate dataset by applying the theorem of Bayes and the concept of conditional independence.

If the so-called Markov properties are fulfilled, and the joint distribution can recursively be factorized into conditional probabilities according to a DAG, it is possible to read conditional independencies off the graph. Since the computation of the joint probability would be intractable in most situations, the factorization helps to simplify computations. Furthermore, the application of graph theory allows to decompose the DAG into smaller subgraphs which again helps to make computations efficient.

Taken together, this tight linkage between probability theory and graphical representation has three advantages (Bishop 2007):

1. A DAG visualizes and defines the structure of a probabilistic model.
2. Conditional independence properties can be read off the graph.
3. Algorithms exploit the graph structure to make inference and model selection in complex graphs possible by breaking down the graph into subgraphs that allow to carry out calculations.

Contrary to the DAGs discussed in the previous sections, probabilistic DAGs do not necessarily require prior knowledge to postulate the DAG. However, they need data for reasoning and decision-making under uncertainty, or when the existence of an edge between two vertices in a DAG has to be estimated. Generally, for interpretation we distinguish the following three purposes of probabilistic networks (Kjærulff and Madsen 2008, p. 18):

1. Deductive reasoning (also referred to as predictive or causal inference) follows the network from the cause to the effect, for example, $P(E \mid O = \text{farmer}, R = \text{rural})$ (see Fig. 39.1).
2. Abductive reasoning (also referred to as diagnostic inference) follows the arrows in reverse order, that is, from the effect to the cause, for example, $P(O \mid E = \text{yes})$.
3. Inter-causal reasoning (also referred to as explaining away) is the ability to make inference on one cause of two independent causes of a known common outcome. This is possible because the outcome is a collider for the two independent causes, which induces an association between the two variables pointing on the collider (see Fig. 39.5). As a result of this induced association, knowing the outcome and having confirmation of one of the causes changes belief in the other cause. This property distinguishes probabilistic networks from other logical systems and is the key feature for making probabilistic networks powerful.

In the next section, we introduce key concepts, formalisms, and some graph terminology, which is required to understand Bayesian networks. Then we discuss how unknown parameters of Bayesian networks and unknown graph structures can be estimated. Beyond this introduction, we recommend the books by Lauritzen (1990), Husmeier (2005), Jensen and Nielsen (2007), Kjærulff and Madsen (2008), Darwiche (2009), and Koller and Friedman (2009) for further reading. Readers who are especially interested in decision-making are referred to the literature on influence diagrams (e.g., Kjærulff and Madsen 2008).

39.3.1 Concepts for Bayesian Networks

For the purpose of a Bayesian network, a directed acyclic graph is defined as follows: A DAG $\mathcal{G}$ is a pair $\mathcal{G} = (V, E)$ with $V = \{1, \ldots, K\}$ the set of distinct vertices and $E \subset (V \times V) \backslash \{(v, u) | v, u \in V, v \neq u\}$ the set of directed edges forming no cycles. Any ordered pair of vertices $(v, u) \in E$ denotes a directed edge and is visualized as arrow $v \rightarrow u$ where v is a parent of u and u is a child of v. The set of all parents and children of a vertex u is given by $pa(u) = \{v \in V | (v, u) \in E\}$ and $ch(u) = \{v \in V | (u, v) \in E\}$, respectively. A directed path of length n from

v_0 to v_n is a sequence of distinct vertices $v_0, \ldots, v_n$ such that $(v_{i-1}, v_i) \in E$ for all $i = 1, \ldots, n$. A directed path is denoted by $v_0 \mapsto v_n$ and a directed cycle is a path of length greater than two and with $v_0 \mapsto v_n$ and $v_n \mapsto v_0$. The vertex u is an ancestor of v and v is descendant of u if there is a directed path $u \mapsto v$. The set $de(u)$ contains all descendants of u and all non-descendants are given as $nd(u) = V \backslash (de(u) \cup \{u\})$. The set of ancestors of u is denoted by $an(u)$ and the ancestral set of u is $An(u) = an(u) \cup \{u\}$. A figure such as $v \mapsto w \mapsfrom u$, where a vertex has unconnected ancestors is called v-structure. For disjoint subsets A, B, S of V, A is separated from B in $\mathcal{G}$ by S if every path between A and B intersects S (see, e.g., Koller and Friedman (2009, p. 35ff.) or Lauritzen (1990, p. 6) for further graph terminology).

Probabilistic DAGs build on the statistical concept of conditional independence and make extensive use of the Bayes' theorem (see chapter ▶Bayesian Methods in Epidemiology of this handbook), also called the chain rule of probability in the context of probabilistic DAGs, where the joint probability of the variables in a DAG is decomposed into products of conditional probability distributions. In the following, let $p(x, y, z)$ denote the probability mass function $P(X = x, Y = y, Z = z)$ in case of discrete random variables and the probability density function if variables are continuous. The chain rule states that the joint probability distribution over a set of variables decomposes into

$$p(x, y, z) = p(z|x, y)p(x, y) = p(z|x, y)p(y|x)p(x).$$

More generally, let us suppose we have K variables. Then the joint probability distribution can be factorized as follows:

$$p(x_1, \ldots, x_K) = \prod_i^K p(x_i|x_{i+1}, \ldots, x_K). \tag{39.1}$$

Example 39.1c.
We assume that *lymphoma* (L), *occupation* (O), and *pesticide exposure* (E) are associated with each other. According to Eq. 39.1, their joint probability distribution can be factorized as

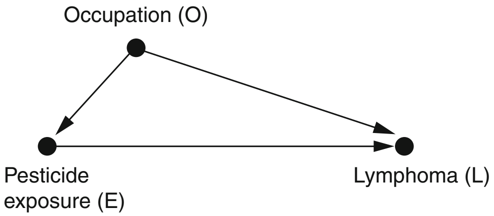
$$p(L, O, E) = p(L|O, E)p(E|O)p(O).$$

The chain rule allows to generate different but interchangeable probability factorizations, as, for example,

$$\begin{aligned} p(L, O, E) &= p(L|O, E)p(O|E)p(E), \\ p(L, O, E) &= p(E|O, L)p(O|L)p(L), \\ &\ldots \end{aligned}$$

Figure 39.6 shows one possibility to construct a DAG representing this joint probability. As we will see later, the DAG in Fig. 39.6 reflects exactly one factorization.

Fig. 39.6 A DAG encoding the joint probability $p(L, O, E)$ between the variables *lymphoma* (L), *occupation* (O), and *pesticide exposure* (E)



The advantage of factorizing the joint probability into a product of conditional probabilities is to make the calculations computationally (more) feasible. A further decomposition is possible if we take advantage of the conditional independencies between the random variables of the random vector $\mathbf{X} = (X_1, \ldots, X_K)'$ (see Box 39.1).

Box 39.1. Conditional independence of random variables

Let X, Y, Z be random variables with joint probability function $p(x, y, z)$. Then X is conditionally independent of Y given Z (written $X \perp\!\!\!\perp Y | Z$) if in the discrete case

$$p(x, y|z) = p(x|z)p(y|z)$$

for all values x, y and for all z such that $p(z) > 0$. If $p(x|z) = p(x)$, $p(z)>0$, then X and Z are marginally independent and we write $X \perp\!\!\!\perp Z$.

The relation $X \perp\!\!\!\perp Y | Z$ means that for every level of Z, Y does not offer any new information to understand X. The conditional independence $X \perp\!\!\!\perp Y | Z$ can also be expressed in slightly different ways as:

$$\begin{aligned} p(x|y, z) &= p(x|z), & p(z) &> 0 \\ p(x, z|y) &= p(x|z)p(z|y), & p(y) > 0, p(z) &> 0. \end{aligned}$$

Example 39.1d.
Suppose in the lymphoma example that we have additional information about the subject's *residence* (R) and we can assume that R is independent from *lymphoma* (L) given information about *occupation* (O) and *pesticide* exposure (E). The joint probability of these four random variables can in general be factorized according to the chain rule as

$$p(L, E, O, R) = p(L|E, O, R)p(E|O, R)p(O|R)p(R)$$

but also as

$$p(L, E, O, R) = p(R|O, E, L)p(O|E, L)p(E|L)p(L).$$

If we make additional use of the conditional independence $R \perp\!\!\!\perp L | \{E, O\}$, the first equation simplifies as follows,

$$p(L, E, O, R) = p(L|E, O)p(E|O, R)p(O|R)p(R).$$

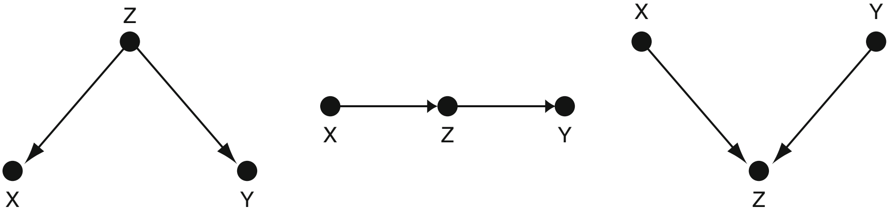


Fig. 39.7 Graphical representation of a fork $(X \perp\!\!\!\perp Y|Z)$, a chain $(X \perp\!\!\!\perp Y|Z)$, and a collider $(X \not\!\perp\!\!\!\perp Y|Z$, but $X \perp\!\!\!\perp Y)$ representing conditional (in)dependence. The fork and the chain represent blocked paths and the collider illustrates an unblocked path

Since conditional independence does not imply any order, it also leads to the following simplification:

$$p(L, E, O, R) = p(R|O, E)p(O|E, L)p(E|L)p(L).$$

These simple rules about factorization and (conditional) independence can be used to reflect (in)dependencies among random variables in a graph as vertices represent random variables, and missing arrows between components of a random vector $\mathbf{X} = (X_1, \ldots, X_K)'$ symbolize (conditional) independencies.

Generally, a conditional independence between three variables X, Y, and Z can be visualized in a DAG in different ways. The fork (see Fig. 39.7) displays a situation where Z regulates X and Y, and hence, X and Y are conditionally independent from each other given Z. The conditional independence $X \perp\!\!\!\perp Y|Z$ can also be represented by a chain in which Z blocks all information flowing from X to Y. In both situations, Z blocks the path between X and Y. If two independent causes X and Y are pointing on a common effect Z, then Z is a collider (see Sect. 39.3.2.2). A collider blocks the path between the marginally independent variables X and Y. However, if we condition on the collider Y, it opens the path and we create an association between X and Y, which implicates the conditional dependence $X \not\!\perp\!\!\!\perp Y|Z$. As Rothman et al. (2008) stated, conditioning on non-colliders closes the path (as the path between X and Y in the chain structure) and conditioning on a collider opens the path (see Sect. 39.2.6).

However, a collider structure also enables to predict values for a parent variable given the knowledge about the other parent and the child. Let us therefore come back to the lymphoma example.

Example 39.1e.
Suppose *occupation* (O) and *residence* (R) are independent causes of *pesticide exposure* (E) as shown in Fig. 39.1. The knowledge about someone working in a chemical plant provides a very good explanation that this person was exposed to pesticides and "explains away" living in a rural region as a possible cause for pesticide exposure. Here, we condition on the collider E, which opens the path between R and O and induces the inequality

$$\begin{aligned} &P(O = \text{chemical industry}|E = \text{yes}, R = \text{rural}) \\ &\qquad\qquad \neq P(O = \text{chemical industry}|E = \text{yes}, R = \text{urban}). \end{aligned}$$

This property is often referred to as "explaining away" effect or "inter-causal inference" (Kjærulff and Madsen 2008). Without conditioning on E, O and R are marginally independent and hence, the information of living in a rural region provides no additional information:

$$P(O = \text{chemical industry}|R = \text{rural}) = P(O = \text{chemical industry}).$$

A further component in the theory of probabilistic DAGs is that the joint probability distribution can be rewritten to combine graph terminology with probability theory by recursively factorizing it as a product of conditional probability distributions (see Box 39.2).

Box 39.2. Recursive factorization

A distribution P allows a recursive factorization according to the structure of a DAG $\mathcal{G} = (V, E)$ if P can be expressed as a product over the local probability distributions of each vertex $v \in V$:

$$p(\mathbf{x}) = \prod_{v \in V} p(x_v|\mathbf{x}_{pa(v)}).$$

Example 39.1f.

The set of parents of the DAGs depicted in Fig. 39.8a are $pa(L) = \{E, O\}, pa(E) = \{O, R\}, pa(O) = \{R\}, pa(R) = \emptyset$ and in Fig. 39.8b $pa(L) = \emptyset, pa(E) = \{L\}, pa(O) = \{E, L\}, pa(R) = \{E, O\}$. Both sets of parents characterize exactly the specific decomposition of their joint probability (see also Example 39.1d). The DAGs illustrated in Fig. 39.8 represent the independence $R \perp\!\!\!\perp L|\{E, O\}$ indicated by the missing edge between R and L. The direction of the arrows between the vertices corresponds to the factorization of the joint density. An arrow points from L to O, that is, $L \rightarrow O$, if $P(L, O)$ factorizes into $P(O|L)P(L)$ and vice versa, that is, $O \rightarrow L$ if $P(L, O) = P(L|O)P(O)$.

So far, we called the link between a DAG and a probability model a probabilistic DAG, whose network structure coincides with a set of conditional probability

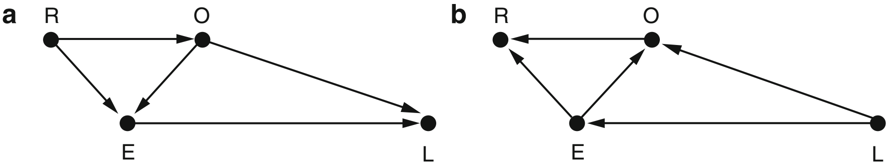


Fig. 39.8 The DAG in (**a**) visualizes the factorization $p(L, E, O, R) = p(L|E, O)p(E|O, R)p(O|R)p(R)$ in contrast to the DAG in (**b**) that represents $p(L, E, O, R) = p(L)p(E|L)p(O|E, L)p(R|O, E)$. Both DAGs encode the conditionally independency $R \perp\!\!\!\perp L|\{E, O\}$

distributions. Since these probabilistic DAGs make use of Bayes' theorem, they are called Bayesian networks (see Box 39.3).

Box 39.3. Bayesian network (cf. Kjærulff and Madsen 2008 p. 65)

A Bayesian network $\mathcal{B} = (\mathbf{X}, \mathcal{G}, P)$ consists of the following:

1. A DAG $\mathcal{G} = (V, E)$ with vertices $V = \{v_1, \ldots, v_K\}$ and directed edges (arrows) E, which describes the dependence structure and implies an ordering of the variables
2. A set of random variables $\mathbf{X} = (X_1, \ldots, X_K)'$, represented by the vertices of $\mathcal{G}$
3. A family of conditional probability distributions P with parameters $\boldsymbol{\theta}$ that admit recursive factorization according to $\mathcal{G}$ where each random variable $X_v \in \mathbf{X}$ is described by one local probability distribution $p(x_v|\mathbf{x}_{pa(v)})$

There are two definitions to assure that the conditional independence statements can be read off the DAG. One of them is the d-separation criterion that is formally defined as in Box 39.4.

Box 39.4. *d*-separation (Pearl 2009, p. 16ff)

A path between v and u is d-separated (or blocked) in a DAG $\mathcal{G} = (V, E)$ by a set of vertices $S \subset V\backslash\{v, u\}$ if and only if at least one of the following conditions hold:

- The path contains a chain $v \mapsto w \mapsto u$ or a fork $v \leftarrow\!\dashv w \mapsto u$ such that $w \in S$, or
- Neither the vertex w nor any descendant of w is an element of S, and w is a collider on the path $v \mapsto w \leftarrow\!\dashv u$.

The sets A and $B \subset V$ are d-separated by $S \subset V\backslash(A \cup B)$ if every path between A and B is blocked by S.

The other definition includes the directed Markov properties (Lauritzen et al. 1990) that require the "moralization" of the graph. The moral graph $\mathcal{G}^m$ is obtained by (i) "marrying," that is, connecting, every pair of non-adjacent parents and then (ii) replacing all arrows with undirected edges. A formal definition of the directed Markov properties is given in Box 39.5.

Box 39.5. Directed Markov properties (Lauritzen 1990, p. 47, 50)

Let $G = (V, E)$ be a DAG and $\mathbf{X} = (X_1, \ldots, X_K)'$ a set of random variables with joint distribution P factorizing recursively according to $\mathcal{G}$ and be $\mathbf{X}_A$ a subvector with components $(X_v, v \in A)$, $A \subset V$. P obeys:

- The *local directed Markov property* if for any vertex $v \in V$

$$X_v \perp\!\!\!\perp \mathbf{X}_{nd(v)\setminus pa(v)} | \mathbf{X}_{pa(v)}$$

- The *global directed Markov property* if for any disjoint sets $A, B, S \subset V$

$$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_S$$

whenever A and B are separated by S in the moral graph $(\mathcal{G}_{An(A \cup B \cup S)})^m$ induced by the smallest ancestral set containing $A \cup B \cup S$

The local and global directed Markov properties are equivalent under the above assumptions.

The d-separation and the directed Markov properties assure that if the joint probability distributions factorize according to the DAG $\mathcal{G}$, then the DAG represents the independence properties of this distribution.

Example 39.1g.
Suppose we want to know whether *residence* (R) and *sex* (S) are conditionally independent given *pesticide exposure* (E) illustrated by the lymphoma DAG in Fig. 39.9a. If we want to read independencies out of the graph, we can apply the global directed Markov property, which requires the moralization of the graph. In the first step, we have to build the DAG that is induced by the ancestral set of R, S, and E, which additionally includes *occupation* (O) as ancestor of E but not *lymphoma* (L), because L is no ancestor of E, R, and S. In the moral graph $(\mathcal{G}_{An(R \cup E \cup S)})^m$ a new edge is needed between *residence* and *occupation* (they have to be "married"), because conditioning on E opens the path between *residence* and *sex* and expresses *residence* and *occupation* as dependent $(R \not\perp\!\!\!\perp S|E)$. It follows that *occupation* has to be additionally included in the separating set modifying the independent statement into $R \perp\!\!\!\perp S|\{E, O\}$. Note that in this example, the subsets R, E, and S contain only one vertex each. However, the concept can be also applied to sets including several vertices. The local directed Markov property gives us the independencies for a specific variable. Suppose we are interested in all independencies regarding *residence*. Applying the above definition, we have to build the set of non-descendants of R without its parents $nd(R)\setminus pa(R) = \{S, O\}\setminus\emptyset = \{S, O\}$. Since the set of parents is empty, we can deduce the following marginal independencies: $R \perp\!\!\!\perp S|\emptyset = R \perp\!\!\!\perp S$ and $R \perp\!\!\!\perp O$.

Each DAG encodes explicit (conditional) independencies that can be translated into the respective factorization of the joint probability using the d-separation or the directed Markov properties. However, different DAGs can induce the same set of conditional independence statements. These DAGs are called Markov-equivalent if and only if they share the same sets of skeletons (graph after replacing all directed edges by undirected edges) and the same sets of v-structures (Andersson et al. 1997; Pearl 2009). Although Markov-equivalent DAGs can look quite different at first glance, they are statistically indistinguishable. Thus, if the aim is to infer the structure of a DAG from given data, many different DAGs may be possible. Even with large amounts of data we cannot determine the "true" structure of the

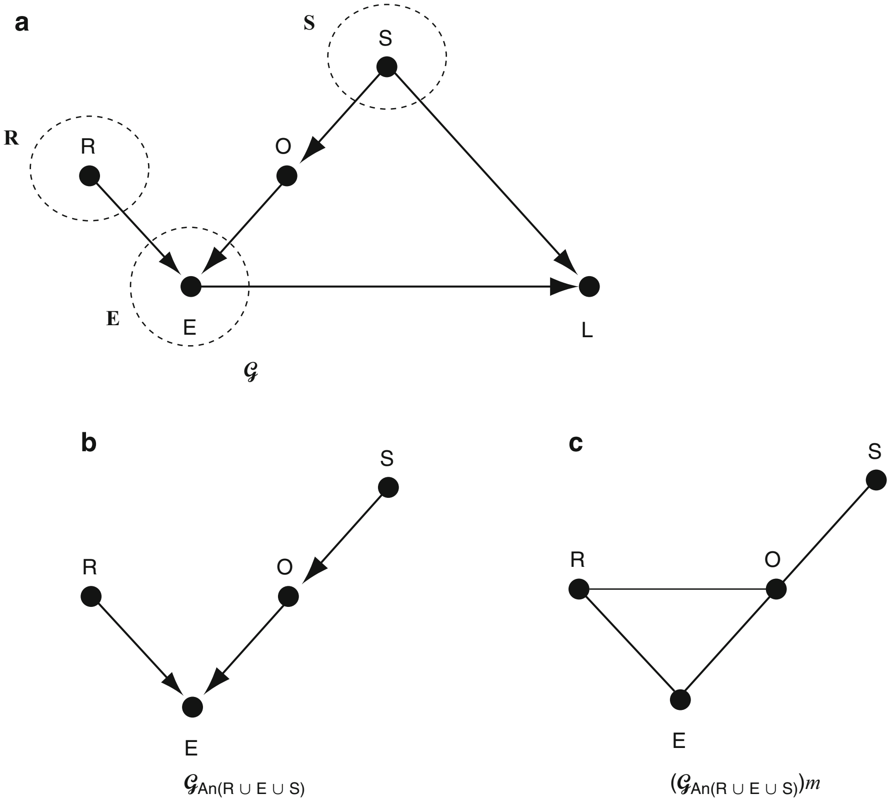

a
S
S
R
R
O
E
E
L
$\mathcal{G}$
b
S
R
O
E
$\mathcal{G}_{An(R \cup E \cup S)}$
c
S
R
O
E
$(\mathcal{G}_{An(R \cup E \cup S)})^m$

Fig. 39.9 (**a**) DAG ($\mathcal{G}$) with subsets R, E, and S; (**b**) DAG induced by the ancestral set of $R \cup E \cup S$; (**c**) the moral graph $(\mathcal{G}_{An(R \cup E \cup S)})^m$

underlying causal graph, but rather (at best) a Markov-equivalent graph. Causal interpretations relying on the direction of an edge can therefore be problematic.

Example 39.1h.
The graphs in Fig. 39.8 share the same skeleton and have both no v-structures. Thus, they both induce the same conditional independence $R \perp\!\!\!\perp L | \{E, O\}$ and are Markov-equivalent, although their arrows point in the opposite direction.

39.3.2 Modeling

The DAG of a Bayesian network is a compact and intuitive representation of uncertain associations between well-defined variables, expressed in (conditional) probabilities. The local probability distribution of each variable $X_v, v \in V$ given its parents defines the conditional distribution as, for example, multinomial, Gaussian, or otherwise. Discrete Bayesian networks contain only discrete variables where the local probability distribution is a set of multinomial distributions parameterized by a probability vector $\boldsymbol{\theta}$ for each configuration of the values corresponding to the parents.

Example 39.1i.

In the lymphoma example, the parents of the binary variable *lymphoma* are *sex* (S), with the states female and male, and *pesticide exposure* (E) with states yes and no. In the framework of a case-control study, the local distributions of L are hence $p(L = \text{case}|S = \text{female}, E = \text{yes})$, $p(L = \text{case}|S = \text{male}, E = \text{yes})$, $p(L = \text{case}|S = \text{female}, E = \text{no})$, and $p(L = \text{case}|S = \text{male}, E = \text{no})$ analogously for $L = \text{control}$ with the same configurations for the parents S and E.

Gaussian Bayesian networks consist only of continuous variables that follow a multivariate normal distribution where the conditional mean depends on the parents' values while the conditional variance does not (Markowetz and Spang 2007). Hybrid Bayesian networks can be applied if discrete and continuous variables have to be modeled simultaneously. These models can be described by the conditional Gaussian distribution (Lauritzen 1992), which assumes that the conditional distribution of the continuous variables given the discrete ones is multivariate Gaussian. Although this distribution allows efficient inference (Cobb et al. 2007), it has the major limitation of restricting the DAG structure too much since discrete vertices cannot have continuous parents. More general models have been presented, including mixtures of truncated exponential models (Moral et al. 2001), mixtures of polynomial models (Shenoy 2011), or Bayesian networks assuming non-parametric distributions (Imoto et al. 2002, 2003). Almost all of these, however, are not implemented in available software yet. Most popular in practice are the multinomial and the multivariate Gaussian distribution.

In contrast to "causal" DAGs, which as discussed earlier are postulated to the best of an epidemiologist's knowledge, Bayesian networks are used when the focus is on (semi-)automatic data-driven modeling. Data analysis applying Bayesian networks may address three different situations:

1. *Expert systems* are fully specified Bayesian networks, including the DAG structure and the parameter vector $\boldsymbol{\theta}$, that is, the probability distribution is completely defined. The aim of expert systems is usually prediction or decision-making under uncertainty in very complex networks. They exploit the graph structure to efficiently compute marginal and conditional probabilities, which otherwise could lead to intractable calculations. Expert systems help to answer questions given some available evidence for one or more variables in the DAG. In the lymphoma example (Fig. 39.1), we could query the likelihood of getting lymphoma for white-collar workers and whether additional information about residence would help to substantially improve the prediction. Expert systems rely on the assumed probabilities and do not necessarily need data, but new data can be used to sequentially update the parameters.

   Lauritzen and Spiegelhalter (1988) developed the probability propagation algorithm to answer those questions in discrete Bayesian networks. Exact inference, however, is an exception and Markov Chain Monte Carlo (MCMC) methods allow for approximate calculations. Expert systems will not be covered in this chapter. For more information, we refer to Cowell et al. (1999), Jensen and Nielsen (2007), or Darwiche (2009).

2. *Parameter learning* aims to estimate the parameters of the local probability distributions. For making inference, we need data and we assume a DAG $\mathcal{G}$ as well

as an underlying distribution with unknown parameters. Recursive factorization over the DAG helps to make these computations feasible in complex systems. Approximate methods for estimating the unknown parameters are needed when missing cases or unmeasured (latent) variables have to be coped with.

3. *Structure learning* aims to find a set of conditional independencies and hence one or more DAGs that describe the data best. For this purpose, a specific distribution, for example, multivariate Gaussian, is assumed without specifying the parameters. In the statistical community, this procedure is referred to as model selection.

Regardless of the aim which Bayesian networks want to achieve, the computations of the conditional probabilities are an exhaustive task due to the large number of parameters in a Bayesian network. A dataset with only 15 binary variables leads to a contingency table with $2^{15} = 32,768$ cells. The calculation of the joint distribution is intractable in such complex situations. Making use of the DAG structure simplifies the complexity since the joint probability factorizes according to the DAG but also because computer algorithms enhance efficiency by exploiting several mathematical properties of these subgraphs for the calculations.

If the dataset is sufficiently large, that is the number of vertices is much smaller than the number of observations, and the dataset does not contain missing values, then exact inference may be possible. Real datasets generally do not fulfill these requirements. Bayesian networks with missing observations (i.e., missing values and unmeasured variables) require approximate estimation techniques where data are expressed as likelihoods rather than as certain values (Borsuk 2008). Computer scientists have developed many algorithms, usually based on optimization methods, to handle missing values and complex association structures. In the following, we will describe inference making and model selection using discrete Bayesian networks in more detail.

39.3.2.1 Parameter Learning: Estimation of the Parameters of a Bayesian Network

This section deals with the estimation of the parameters of the local distributions of a Bayesian network. Parameter learning requires two assumptions: firstly, the structure of the DAG is known, and hence, parameter learning does not aim to estimate the direction of an arrow between two variables and secondly, a complete random sample $\mathcal{D} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ from the joint probability distribution of $\mathbf{X}$ is given.

If the dataset comprises enough data, we can use the principle of maximum likelihood to infer the parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)'$. Let $\mathcal{B} = (\mathbf{X}, \mathcal{G}, P)$ be a Bayesian network which is characterized by its graph structure $\mathcal{G}$ and a set P of local probability distributions associated with each variable, and let $\mathcal{D}$ be a dataset of complete observations. The likelihood $L(\boldsymbol{\theta}, \mathcal{G}|\mathcal{D}) = p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{G})$ is maximized with respect to the parameters $\boldsymbol{\theta}$ of the local probability distribution $p(x_v|\mathbf{x}_{pa(v)})$ exploiting the DAG structure $\mathcal{G}$ to obtain the maximum likelihood (ML) estimator as

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\arg\max}\, L(\boldsymbol{\theta}, \mathcal{G}|\mathcal{D}).$$

Let us assume the data are discrete to demonstrate the basic idea. Each variable $X_v \in \mathbf{X}$ has r_v possible categories and each local distribution is a set of multinomial distributions, one multinomial distribution for each parent configuration according to $\mathcal{G}$:

$$p(X_v = j|\mathbf{X}_{pa(v)} = l, \theta_v, \mathcal{G}) = \theta_{vjl} > 0,$$

where $\theta_v = \left(\left(\theta_{vjl}\right)_{j=1}^{r_v}\right)_{l=1}^{q_v} = (\theta_{v11}, \ldots, \theta_{vr_K q_K})$ are the parameters under the constraint $\sum_{j=1}^{r_v} \theta_{vjl} = 1$ with q_v denoting the number of possible parent configurations of vertex v.

The recursive factorization property (see Box 39.2) is applied to decompose the likelihood of a Bayesian network to get K independent and less complex estimation problems:

$$L(\boldsymbol{\theta}, \mathcal{G}|\mathcal{D}) = p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{G}) = \prod_{v=1}^{K} p(x_v|\mathbf{x}_{pa(v)}, \theta_v, \mathcal{G}) = \prod_{v=1}^{K} L(\theta_v, \mathcal{G}|\mathcal{D}).$$

The ML estimator of $\theta_{vjl} = p(X_v = j|\mathbf{X}_{pa(v)} = l, \theta_v, \mathcal{G})$ for multinomially distributed data is given as

$$\hat{\theta}_{vjl} = \frac{N_{vjl}}{\sum_{l=1}^{q_v} N_{vjl}} = \frac{\sharp \text{ subjects with } X_v = j \text{ and } \mathbf{X}_{pa(v)} = l}{\sharp \text{ subjects with } \mathbf{X}_{pa(v)} = l}. \tag{39.2}$$

Example 39.2a.
Consider a collider structure (Fig. 39.10) corresponding to a mechanism where genetic (G) and environmental (E) factors independently influence a disease (Y). In this setting, the genotype of the genetic factor G is considered as either "common" or "rare," and a person is either "exposed" or "unexposed" to the environmental factor E. We want to calculate the prevalence of Y given the ML estimates of the multinomial local probability distributions. Table 39.1 shows the ML estimators $\hat{\theta}_{vjl}$ for each vertex in the DAG obtained from a sample with $N = 2,307$ observations. We omit the results for the state "healthy" of the

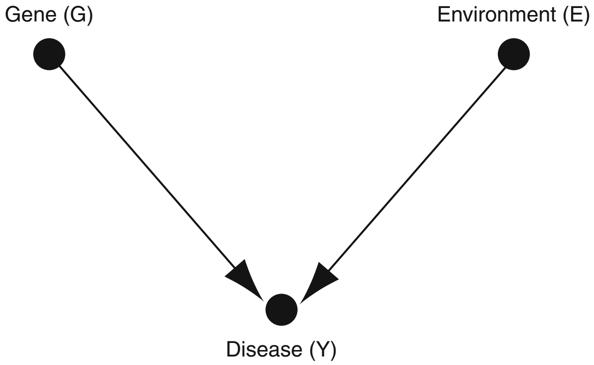


Fig. 39.10 Graphical representation of Example 39.2a

Table 39.1 ML estimates for the local probability distributions of G, E, and Y using Eq. 39.2. The ML estimators are based on the value j of the vertex v and the combination l of the values of v's parents. The vertices G and E have no parents and hence, their ML estimator simplifies to $\hat{\theta}_{vj} = \frac{N_{vj}}{N}$

v	j	l	N_{vj}	$\hat{\theta}_{vj}$
G	Common	–	2, 147	0.931
	Rare	–	160	0.069
E	Unexposed	–	2, 091	0.887
	Exposed	–	261	0.113
			N_{vjl}	$\hat{\theta}_{vjl}$
Y	Diseased	Common – unexposed	13	0.131
	Diseased	Common – exposed	128	0.269
	Diseased	Rare – unexposed	7	0.045
	Diseased	Rare – exposed	8	0.008

variable Y. G and E have no parents, and the parent configuration of Y are the four possible combinations of the values common/rare and exposed/unexposed.

The estimator $\hat{p}(Y =$"diseased") is computed using recursive factorization according to the DAG structure as

$$\begin{aligned}\hat{p}(Y) &= \sum_g \sum_e \hat{p}(Y = y|G = g, E = e)\hat{p}(G = g)\hat{p}(E = e)\\ &= (0.131 \cdot 0.069 \cdot 0.113) + (0.269 \cdot 0.931 \cdot 0.113)\\ &\quad +(0.045 \cdot 0.069 \cdot 0.887) + (0.008 \cdot 0.931 \cdot 0.887)\\ &= 0.00103 + 0.02837 + 0.00276 + 0.00664 = 0.03880\end{aligned}$$

which yields $\hat{p}(Y =$ "diseased"$) = 0.0388$, that is, the prevalence is about 4%.

ML estimation has the disadvantage of being sensitive to small sample sizes. Consider a problem, where we have sparse data and many discrete variables that can be arranged in a large contingency table, which will probably have many cells with zero entries. In the discrete example, ML estimation relies on frequency counts and will therefore give cells with zero counts also a probability of zero, which has a (misleading) strong influence on the outcome. An alternative to ML estimation is Bayesian inference that assumes $\boldsymbol{\theta}$ also as unknown, yet not fixed. Bayesian estimation treats $\boldsymbol{\theta}$ as a random vector and takes prior knowledge about each parameter into account to estimate the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{G})$.

Let us consider a given DAG $\mathcal{G}$ and data with likelihood $L(\boldsymbol{\theta}, \mathcal{G}|\mathcal{D})$. Additionally assume a prior distribution $p(\boldsymbol{\theta}|\mathcal{G})$ of the parameters $\boldsymbol{\theta}$ for a given graph $\mathcal{G}$ to derive the posterior distribution of the parameters

$$p(\boldsymbol{\theta}|\mathcal{G}, \mathcal{D}) = \frac{p(\boldsymbol{\theta}|\mathcal{G})p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{G})}{p(\mathcal{D}|\mathcal{G})} \propto p(\boldsymbol{\theta}|\mathcal{G})L(\boldsymbol{\theta}, \mathcal{G}|\mathcal{D}).$$

Applying Bayes' theorem, the posterior is proportional to the product of the prior distribution and the likelihood. The normalizing denominator $p(\mathcal{D}|\mathcal{G})$ reflects the average probability of the data under a given DAG, but it does not depend on $\boldsymbol{\theta}$, and thus its computation is not of interest. The influence of the prior information on the posterior distribution will decrease with increasing sample size, that is, when the evidence from the data overrules the prior information or if the prior information is uninformative (see chapter ▶Bayesian Methods in Epidemiology of this handbook). A proper factorization of the posterior distribution depends on the choice of the distribution for the priors $p(\boldsymbol{\theta}|\mathcal{G})$. A suitable prior is the conjugate prior, that is, when the posterior distribution is of the same family as the prior distribution. The conjugate prior for multinomial data is the Dirichlet distribution. If no conjugate prior distribution can be found, it is impossible to compute the posterior exactly, and the posterior distribution has to be approximated.

To make computations generally feasible, that is, also for large networks, it is desirable that the parameters do not have to be learned simultaneously but rather that each component θ_{vjl} of the parameter vector $\boldsymbol{\theta}$ can be learned individually, which needs the assumption of no missing data in the random sample $\mathcal{D}$ and the assumption of parameter independence (Spiegelhalter and Lauritzen 1990). This means that (a) the parameters θ_v of each local distribution are independent of each other (global independence), allowing the decomposition $p(\boldsymbol{\theta}|\mathcal{G}) = \prod_{v=1}^{K} p(\theta_v|\mathcal{G})$, and (b) that the parameters θ_v are independent for each parent configuration $\mathbf{x}_{pa(v)}$ (local independence), meaning for each variable v it follows $p(\theta_v|\mathcal{G}) = \prod_{j=1}^{r_v} \prod_{l=1}^{q_v} p(\theta_{vjl}|\mathcal{G})$. If global and local independence hold, then the joint distribution factorizes as follows:

$$p(\boldsymbol{\theta}|\mathcal{G}) = \prod_{v=1}^{K} \prod_{j=1}^{r_v} \prod_{l=1}^{q_v} p(\theta_{vjl}|\mathcal{G}).$$

Example 39.2b.
Figure 39.11 shows a Bayesian network representation for Example 39.2a using a Bayesian approach. Nodes for the prior distributions are placed outside the gray box and parameter independence holds since they are all independent from each other. The observations $X^1, \ldots, X^N$ are conditionally independent given θ_v.

Up to now we assumed a dataset without any missing data points or unobserved variables. Real data, however, are frequently incomplete, so that some variables show missing values or certain variables remain unmeasured. In such situations, the posterior distribution factorizes into components that are not well known anymore, which leads to analytically intractable calculations and high computational complexity. Hence, we need approximative and more complex solutions in order to make inference.

ML estimation with data missing at random (see chapter ▶Missing Data of this handbook) is generally achieved by the Expectation-Maximization (EM) algorithm (see Lauritzen (1995) for Bayesian networks). Applying the idea of Bayesian

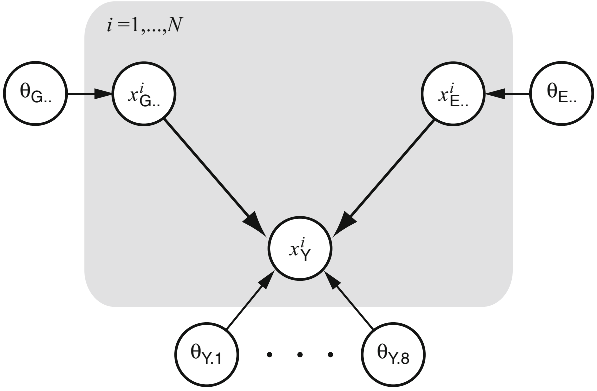


Fig. 39.11 Graphical representation of Example 39.2a using a Bayesian approach under the assumption of parameter independence. Parameters are unknown and marginally independent having the joint probability $p(\mathbf{x}, \boldsymbol{\theta} | \mathcal{G}) = \prod_{v \in V} p(\theta_v | \mathcal{G}) \prod_{i=1}^{N} p(x^i_v | \mathbf{x}^i_{pa(v)}, \theta_v, \mathcal{G})$

inference, the most popular approaches for approximative inference are based on MCMC techniques. An example is the Gibbs sampler (e.g., Gilks et al. 1996).

Since with increasing sample size the influence of the prior decreases and the posterior approaches the likehood, frequentist and Bayesian inference are asymptotically equivalent when estimating unknown parameters. However, Bayesian approaches are generally less computationally expensive than frequentist approaches (Husmeier 2005, p. 26). Beyond that, Chickering (1996), Chickering et al. (2004), and Dagum and Luby (1993) have shown that exact as well as asymptotic probabilistic inference in Bayesian networks is NP-hard (non-deterministic polynomial-time hard). This means that no general algorithm exists that can perform a solution to this problem in less than exponential time (exponential with respect to the number of variables in the DAG). Suitable algorithms are therefore those that can reduce the required (possibly exponential) time and present even so adequate estimates. See Daly et al. (2011) for an overview.

39.3.2.2 Structure Learning: Model Selection of a Bayesian Network

In epidemiology we are often confronted with situations where we neither know the “true” causal graph structure of the target population nor the parameters that describe this structure. The aim is then to learn a Bayesian network from data where the graph and the parameters are unknown. The selected DAG with its probabilistic model or learned Bayesian network should explain the association structure among a given set of random variables and data $\mathcal{D}$ as best as possible.

A naïve optimization method to identify the DAG with the best fit would be to compare the fit of every possible DAG and then select the Bayesian network with the best fit. This intuitive model search strategy is too exhaustive as the number of possible DAGs g_K is super-exponential in the number of vertices K in the DAG.

The number of possible DAGs can be calculated with Robinson's recursive formulae (Robinson 1977) as

$$g_K = \sum_{v=1}^{K} (-1)^{v-1} \binom{K}{v} 2^{v(K-v)} g_{K-v} \quad \text{with } g_0 = 1$$

which points out that exhaustive search is infeasible. In Example 39.1a with five vertices, we would have to calculate $29,281$ networks; ten vertices would already imply approximately $4.8 \cdot 10^{18}$ DAGs. The goal is hence to identify one or more DAGs with a good fit in acceptable time.

Two types of search strategies are typically distinguished to achieve this: (i) score-based algorithms that follow a scoring approach to differentiate between several Bayesian networks and (ii) constraint-based algorithms that apply tests of independence to decide about acceptance of an edge. Sometimes this strategy tries to follow causal semantics to select a DAG. There are also hybrid search strategies that combine score-based and constraint-based algorithms. Once the DAG is selected, the parameters of the Bayesian network can be estimated using, for example, the methods mentioned in Sect. 39.3.2.1.

Score-Based Algorithms Score-based algorithms usually consist of two components. The first component involves a scoring criterion that measures the degree to which a selected graph fits the data given prior knowledge (if any) and assigns a value to that DAG. The second component is a search strategy to identify DAGs with high scoring values.

Using the method of maximum likelihood, the best fit for a given graph is achieved by maximizing the likelihood $L(\boldsymbol{\theta}_{\mathcal{G}}, \mathcal{G}|\mathcal{D})$ of its parameters $\boldsymbol{\theta}_{\mathcal{G}}$ given the data $\mathcal{D}$. An ML score for a graph $\mathcal{G}$ is then given as

$$\text{ML}_{\text{score}} = \max_{\boldsymbol{\theta}_{\mathcal{G}}} \log L(\boldsymbol{\theta}_{\mathcal{G}}, \mathcal{G}|\mathcal{D}) = \log L(\hat{\boldsymbol{\theta}}_{\mathcal{G}}, \mathcal{G}|\mathcal{D}).$$

The ML score usually suffers from overfitting as the score increases with the number of parameters in the model. Generally, it will judge the saturated model (each vertex is a parent of all its non-parents) as the best fit, which is useless. Most scoring criteria combine therefore model fit and model complexity to avoid overfitting. One example is the Bayesian information criterion (BIC). It determines the model fit based on the ML method from above and controls overfitting by penalizing the maximum likelihood by the number of model parameters d:

$$\text{BIC} = \log L(\hat{\boldsymbol{\theta}}_{\mathcal{G}}, \mathcal{G}|\mathcal{D}) - \tfrac{d}{2} \log N,$$

where the sample size N calibrates the penalty.

Another ML-based scoring criterion is the Akaike information criterion (AIC) that penalizes the ML estimator only by the number of parameters.

The BIC is a consistent scoring criterion (Geiger et al. 2001; Chickering and Meek 2002). This means that under the assumption that the true model is one of the candidate models, the BIC will assign the true model the highest score with probability one for N tending to infinity. The AIC is not consistent but is an efficient scoring criterion. The AIC is a good choice for high-dimensional settings where the true causal graph is not believed to be in the set of candidate models and the aim is to adequately describe the unknown DAG. Graphs selected by the AIC are typically more complex, that is, the DAG contains more edges than those selected by the BIC, which in turn tends to be more parsimonious, yet in smaller graphs also too parsimonious. Although motivated from a Bayesian point of view, the BIC does not reflect a Bayesian approach since it does not depend on a prior distribution for the parameters.

The Bayesian score (BS) is an often preferred alternative to frequentist methods. Suppose that $\boldsymbol{\theta}_{\mathcal{G}}$ and $\mathcal{G}$ are random variables with the structural prior $p(\mathcal{G})$ and parameter prior $p(\boldsymbol{\theta}_{\mathcal{G}}|\mathcal{G})$. The posterior distribution of a graph $\mathcal{G}$ given data $\mathcal{D}$ is then

$$p(\boldsymbol{\theta}_{\mathcal{G}}, \mathcal{G}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{G})}{p(\mathcal{D})},$$

where $p(\mathcal{D}|\mathcal{G}) = \int p(\mathcal{D}|\boldsymbol{\theta}_{\mathcal{G}}, \mathcal{G})p(\boldsymbol{\theta}_{\mathcal{G}}|\mathcal{G})d\boldsymbol{\theta}_{\mathcal{G}}$ is called the marginal likelihood of the graph $\mathcal{G}$. It expresses the averaged likelihood over all possible parameters of the local distributions. The parameters are marginalized out of $p(\mathcal{D}|\mathcal{G})$ by integrating the marginal likelihood with respect to $\boldsymbol{\theta}_{\mathcal{G}}$. The denominator $p(\mathcal{D})$ neither depends on $\mathcal{G}$ nor on $\boldsymbol{\theta}_{\mathcal{G}}$ and can be treated as constant which is the same for all models and has therefore not to be computed for model selection. The marginal likelihood can be computed under reasonable assumptions, that is, assuming a random sample from the unknown data-generating causal graph with no missing values and no missing variables, parameter independence, and a conjugate prior. The Bayesian score of $\mathcal{G}$ is defined as

$$\text{BS} = \log p(\mathcal{G}, \mathcal{D}) = \log p(\mathcal{D}|\mathcal{G}) + \log p(\mathcal{G}),$$

which avoids overfitting by averaging over the parameters $\boldsymbol{\theta}_{\mathcal{G}}$ and not maximizing the likelihood.

The structural prior $p(\mathcal{G})$ is mostly chosen as uniform (each DAG has equal probability). However, it is also possible to choose a prior distribution to account for biological knowledge by excluding graphs or assigning them different probabilities. It should be noted that sometimes the optimal model can only be detected by searching through some impossible models. With increasing sample size, $p(\mathcal{G})$ plays a minor role as it remains constant while the marginal likelihood grows linearly.

Different scoring criteria exist to compute the posterior probability. The K2 scoring function (Cooper and Herskovits 1992) is, for example, a possible choice if among others a Dirichlet distribution is assumed as conjugate prior for multinomial data as well as parameter independence and complete data.

The score was further developed by Heckerman et al. (1995) to the Bayesian Dirichlet equivalence (BDe) metric, which is score equivalent in contrast to K2. Score equivalence means that Markov-equivalent DAGs will be assigned the same score value.

The accuracy of a model selected based on the Bayesian score criterion is higher than if it is selected based on the BIC. However, the model selection procedure using the Bayesian score is also more computationally expensive. Liu et al. (2012) empirically evaluated different scoring criteria for Bayesian network model selection and concluded that the BIC consistently outperforms other scoring functions.

The choice of the search space (e.g., network structures, equivalence classes of network structures), a proper scoring criterion, and a strategy how to search through the space of graph structures are important. Score-based algorithms use one of the above-mentioned scoring criteria to rank different models by maximizing the scoring function over some search space to identify the Bayesian network with the highest score.

A popular search strategy is heuristic greedy search. It aims to find a satisfactory solution in reasonable time by choosing a graph that offers an immediate improvement without considering future consequences. This approach does not identify the global optimal graph, that is, the DAG with the highest score of all possible DAGs, but rather finds a local optimal graph meaning that each modification of one (or just a few) edge promises no higher score.

One well-known score-based greedy algorithm is the hill-climbing algorithm. It starts with an initial graph and modifies it stepwise by adding, deleting, or reversing an arrow until the score of the final model cannot be improved. Hill climbing has the disadvantage that it may stick at a plateau or at a local maximum and does not find the global maximum. As explained before, all neighboring graphs (graphs with one modified arrow) show either the same or a lower score, although the current model score is not the highest of all possible models. Solutions to avoid local maxima and plateaus include simulated annealing (Kirkpatrick et al. 1983; Darwiche 2009) or the repetition of the model selection with random initial models. Another possibility is to use tabu search (Glover 1989, 1990) instead, which enhances the performance by memorizing visited networks.

Some algorithms reduce the search space by searching among Markov-equivalent classes of Bayesian networks. DAGs within a Markov-equivalent class determine the same probability model and will therefore be assigned the same score. Several researchers (Chickering 1996; Andersson et al. 1997) prefer to search in the space of equivalence classes rather than in the space of all DAGs.

Fundamental general techniques of structure learning algorithms were developed by Heckerman et al. (1995) and Chickering (1996) and further improved (see Markowetz and Spang (2007), Daly et al. (2011), and Koller and Friedman (2009) for more information).

Constraint-Based Algorithms The aim of constraint-based approaches is to determine DAGs that entail the directed Markov properties (or equivalently the d-separation relations) corresponding to the independencies among the components of the random vector $\mathbf{X}$. A sequence of (conditional) independence tests is applied to establish a set of (conditional) independencies that are used to construct the graph edge by edge. Algorithms first form an undirected version of the DAG before the directions of the arrows are included in the next step.

Some algorithms optimize their selection strategy by limiting the search space to the *Markov blanket* of each vertex. The Markov blanket of a vertex v contains the parents, children, and all vertices that share a child with v (Scutari 2010). The final graph then specifies the constraints of the Markov property for the data. Again, heuristic algorithms are necessary since exhaustive search is computationally infeasible.

Statistical hypothesis tests detecting (conditional) independence for discrete data include association tests like Pearson's χ^2-test or deviance-based tests. Conditional independence tests for continuous variables are based on the partial correlation coefficient of X and Y given Z. For many tests, exact and asymptotic versions are available.

Some algorithms search for "causal" DAGs. The inductive causation (IC) algorithm by Verma and Pearl (1991, 1992) is an intuitive example. The procedure can roughly be summarized as follows (Jensen and Nielsen 2007, p. 231ff.):

1. Find the *conditional independencies* and learn the skeleton of the graph, that is, the graph contains only undirected edges.
2. *Identify all v-structures* For triplet-structures $X - Y - Z$, check if Y is in the separating set S of $X \perp\!\!\!\perp Z | S$. If $Y \in S$ then X and Z are conditionally independent given Y. If $Y \notin S$ then introduce the v-structure $X \rightarrow Y \leftarrow Z$ in the graph.
3. *Avoid new v-structures* If there are missing directions in a directed path, like $X \rightarrow Y - Z$ and Z is not adjacent to X, then orient $X \rightarrow Y \rightarrow Z$.
4. *Avoid cycles* If $X \rightarrow Y$ introduces a directed cycle in the graph, then reverse the arrow $X \leftarrow Y$.
5. *Choose direction randomly* (optional) If none of Steps 2–4 can be applied, choose an undirected edge and direct it randomly.

Example 39.1h.
Figure 39.12a shows the skeleton of the lymphoma example with the found (conditional) independencies. Applying Steps 2–4 of the IC algorithm results in the graph in Figure 39.12b. The graph containing the edge $O \rightarrow S$ instead of $O \leftarrow S$, however, is Markov-equivalent.

Although the IC algorithm is simple to understand, it is not specific enough for practical application. The first successful implementation of a causal discovery algorithm was developed by Spirtes and Glymour (1990). Their PC algorithm relies on conditional independence tests to build a preliminary undirected graph and a systematic procedure to remove unnecessary edges and adjust directions by applying the directed Markov properties or the d-separation. The output of the PC algorithm is an acyclic, partially directed graph, that is, some edges are undirected. DAGs with undirected edges represent an equivalence class, where all DAGs within that equivalence class are Markov-equivalent and represent the same set of independence properties (see Fig. 39.13). The PC algorithm is consistent if the distribution P is faithful with respect to a graph $\mathcal{G}$, meaning that the conditional independencies of the distribution can be transferred to $\mathcal{G}$ ($\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_S \Rightarrow A \perp\!\!\!\perp B | S$), and all conditional independencies have been correctly identified. Under certain assumptions, the PC algorithm is a consistent estimator for causal effects,

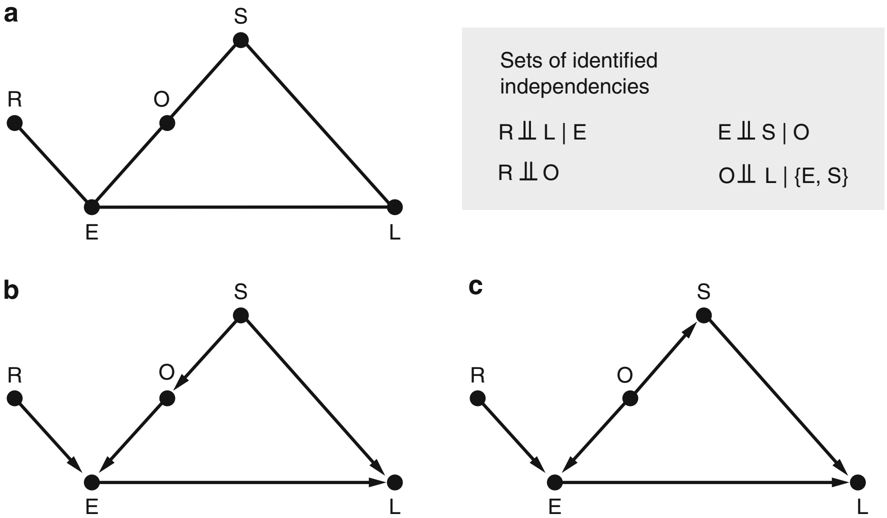


Fig. 39.12 (**a**) Skeleton encoding the (conditional) independencies; (**b**) graph after introducing all v-structures and chain structures, according to the IC algorithm; (**c**) a Markov-equivalent graph

although there might be unobserved confounders and/or unknown time order. Robins et al. (2003) showed that the type of consistency is not uniform, implying that the PC algorithm selects the true graph with an infinite sample size. That means, even in very large finite samples, the true graph will not necessarily be selected.

There are several modifications of the PC algorithm. The PC* algorithm significantly reduces the number of performed independence tests and is therefore quicker than the PC algorithm (Spirtes et al. 2001, pp. 84–89). The CPC algorithm (Ramsey et al. 2006) uses a more sensitive test for orientation of colliders, and the JPC algorithm (Ramsey 2010) is restricted to continuous variables which tries to improve accuracy. Kalisch and Bühlmann (2007) have shown that assuming Gaussian data, the PC algorithm is applicable to high-dimensional data.

It is important to remember that with data-driven model selection based upon conditional independence tests, it is only possible to construct a DAG up to its Markov equivalence class (Kalisch et al. 2012). To direct undirected edges is then only possible with prior knowledge about causal directions.

Some limitations of statistical hypothesis tests also need to be considered in this context. An independence test is constructed to prove statistical dependence. However, statistical independence is accepted when the data shows no evidence against independence. This does not mean that the true data-generating process implies independence. Additionally, each statistical test has a probability for false-positive or false-negative decisions. False decisions have the consequence that edges in the DAG are erroneously added or deleted. The selected DAG then represents independence statements that are at least partly untrue. The accuracy of

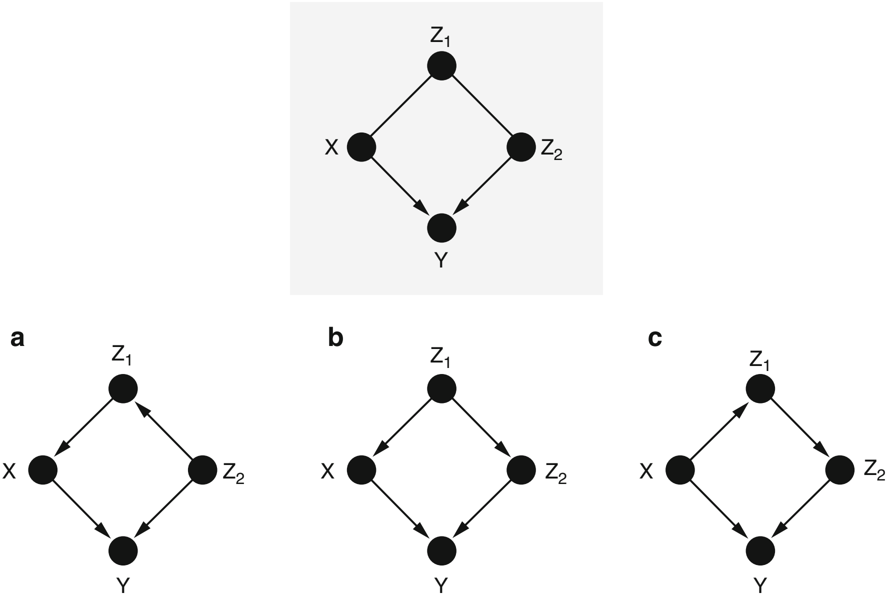


Fig. 39.13 The graph in the *gray box* illustrates the Markov equivalence class of the three Markov-equivalent graphs (**a**)–(**c**)

constraint-based methods is discussed by Fast et al. (2008) in a more general setting. Li and Wang (2009) developed a procedure to control the false discovery rate of the learned graph using the PC algorithm.

Score-based and constraint-based methods have each their advantages. Constraint-based algorithms are more efficient when the graph is sparse. They can deal better with latent variables and even work in the presence of selection bias. However, the most serious disadvantage is their insufficiency to adequately cope with the problem of simultaneously testing multiple hypotheses. Score-based techniques, on the one hand, show a better performance, especially for dense graphs and with less data, they favor parsimonious models and thus avoid overfitting, and they are superior in dealing with missing data. On the other hand, they "only" find a good model but do not identify the best one. Hybrid algorithms have been developed to combine the advantages of both methods (e.g., Wong et al. 2002; Tsamardinos et al. 2006; Wang et al. 2007). In a first phase, the hybrid procedures usually apply a constraint-based algorithm to learn an initial skeleton. In a second phase, the score-based algorithm searches in the reduced search space for the Bayesian network with the highest score. The inclusion of latent variables or the consideration of selection bias in the network requires specific algorithms (Friedman 1997; Spirtes et al. 1995).

Our discussion about Bayesian networks is far from comprehensive. Beside the monographs mentioned earlier, the tutorial by Heckerman (1999) represents a good introduction into the topic, especially on parameter learning. In Daly et al. (2011), a good review of existing algorithms for learning parameters and structure can be found. Darwiche (2010) gave a broad overview of applications of Bayesian networks. Issues of robustness and sensitivity are, discussed, for example, by Kjærulff and Madsen (2008, Chap. 10), Husmeier (2003), and Friedman et al. (1999b, a). Genetic and biological aspects are covered in the reviews by, for example, Friedman (2004) or Markowetz and Spang (2007). Applications in epidemiology outside genetic epidemiology are rare to date (Getoor et al. 2004; Nguefack-Tsague 2011; Nadathur and Warren 2011; Stefanini et al. 2009).

39.3.3 Software

As the previous discussions show, a major issue in working with Bayesian networks is the choice and availability of appropriate software. Numerous software applications are available (Korb and Nicholson 2011; Murphy 2007) which we describe roughly. `R` (R Development Core Team 2012) offers several packages to deal with Bayesian networks. One of them is the `gRbase` package (Dethlefsen and Højsgaard 2005), which is a platform for a more general framework of graphical models. `gRain` (Højsgaard 2012) uses this platform to compute expert systems. The book by Højsgaard et al. (2012) introduces graphical modeling using `R` but focuses mostly on `gR`-packages. The package `bnlearn` (Scutari 2010) includes several algorithms for learning Bayesian networks with either discrete or continuous data and provides parallel computing. The package `pcalg` (Kalisch et al. 2012) implements the PC algorithm and other constraint-based algorithms. Bayesian networks with mixed variables can be handled by `deal` (Bottcher and Dethlefsen 2011). Popular standalone software packages include the commercial software `Hugin` for building and making inference from expert systems (Madsen et al. 2003), OpenBugs for parameter learning using MCMC methods (Lunn et al. 2009), or `Tetrad IV` (Glymour et al. 2012) for building and calculating Bayesian networks applying the PC algorithm and many other algorithms described in Spirtes et al. (2001). An updated, but not complete list of software packages for Bayesian networks is maintained by Murphy (2012).

39.4 Conclusions

In epidemiology, DAGs are typically applied as a tool to visualize an etiological network. Ideally, this visualization allows to efficiently identify potential confounders that have to be adjusted for in the estimation of the effect of a risk factor on the outcome under study. DAGs are usually constructed on the basis of single hypotheses since the entire etiological network is usually unknown. The construction of a DAG therefore requires a comprehensive understanding of the etiology to develop the optimal DAG. However, when not data-based, the construction of a DAG relies on assumptions which are always incomplete and may lead to different DAGs. In situations where edges or directions are uncertain, Shrier and Platt (2008) recommend to postulate alternative, equally plausible DAGs for the situation under study. These alternative DAGs should be tested to evaluate whether they generate the same adjustment set of covariates, and results of these alternative DAGs should also be published. Such sensitivity analyses help to assess the robustness of the estimated effect measures (Geneletti et al. 2011).

A considerable limitation of DAGs, if only considered as a graph and not as a probability model, is the fact that the arrows in the graph itself do not indicate the type (protective or risk-enhancing) and strength of the associations. While the (likely) presence of bias is identifiable from a well-developed graph, no information on the extent and direction of the bias and its relative importance in mathematical terms can be derived from the DAG alone. The use of Bayesian networks can be helpful in this respect as estimation of parameters is possible here. So far, DAGs cannot represent interactions or effect modification (Hernán et al. 2004), and some authors are very skeptical whether DAGs will ever provide a reasonable solution to disentangle the complexity of causal networks (Weinberg 2007). In spite of this criticism, VanderWeele and Robins (2007a, b) have started to deal with effect modification in DAGs, but their models require stringent assumptions.

The main difference between a causal DAG and a probabilistic DAG is the fact that the construction of the causal DAG relies on adding an arrow $X \rightarrow Y$ whenever we think about a causal (dependent) flow from X to Y, while the second approach is built upon (conditional) independencies. This admittedly is a somewhat complicated way of thinking. It reveals the association patterns embedded in the data particularly when no reliable background knowledge exists. A further difference between causal and probabilistic DAGs is the fact that cause-effect relationships are non-symmetrical. Arrows indicate the direction of an association, whereas probabilistic independence associations are symmetrical and the connection between two vertices symbolizes "no (conditional) independence" between two vertices without implying a direction. Moreover, from a statistical point of view, it is a challenge to distinguish association from causality. DAGs have become increasingly popular to postulate causal relationships, but data-driven modeling of DAGs can at best assist to explore causal dependence networks. This results from the fact that, so far, a single DAG cannot accomplish probabilistic and causal interpretations simultaneously (Dawid 2010a). For example, we showed earlier that Markov-equivalent DAGs encode for the same conditional probabilities but have different "causal" interpretations. Markov-equivalent graphs can therefore not be distinguished based on the data alone.

An overall problem is that probability distributions are well defined, whereas the concept of causality has no common understanding in epidemiology. Therefore, how to draw causal conclusions from non-experimental data remains a key challenge for epidemiologists, especially when confounding is not fully understood. As a consequence, prominent researchers in this field (Pearl 2009; Spirtes et al. 2001; Dawid 2010b; Didelez and Sheehan 2007) have been arguing for a formal language for causality. Pearl's *do*-operator (Pearl 2009) or counterfactuals (Rubin 1974) are examples of this formalization. Here, causality is generally viewed as an intervention effect in a given system, and DAGs help to clarify the assumptions of these concepts. A formal language is also necessary to improve and invent causal algorithms for causal discovery using Bayesian networks.

References

Andersson SA, Madigan D, Perlman MD (1997) A characterization of Markov equivalence classes for acyclic digraphs. Ann Stat 25:505–541

Berkson J (1946) Limitations of the application of fourfold tables to hospital data. Biom Bull 2:47–53

Bishop CM (2007) Pattern recognition and machine learning. Springer, New York

Borsuk ME (2008) Bayesian networks. In: Jørgensen SE, Fath B (eds) Encyclopedia of ecology. Elsevier, Burlington, pp 307–317

Bottcher SG, Dethlefsen C (2011) Deal: learning bayesian networks with mixed variables. http://CRAN.R-project.org/package=deal. R package version 1.2–34

Breitling L (2010) dagR: a suite of R functions for directed acyclic graphs. Epidemiology 21: 586–587

Chickering D, Meek C (2002) Finding optimal Bayesian networks. In: Darwiche A, Friedman N (eds) Proceedings of the eighteenth annual conference on uncertainty in artificial intelligence (UAI-02). Morgan Kaufmann, San Francisco, pp 94–102

Chickering DM (1996) Learning Bayesian networks is NP-complete. In: Fisher D, Lenz HJ (eds) Learning from data: artificial intelligence and statistics V. Lecture notes in statistics, vol 112. Springer, New York, pp 121–130

Chickering DM, Heckerman D, Meek C (2004) Large-sample learning of Bayesian networks is NP-hard. J Mach Learn Res 5:1287–1330

Cobb BR, Rumí R, Salmerón A (2007) Bayesian network models with discrete and continuous variables. In: Lucas P, Gámez JA, Salmerón A (eds) Advances in probabilistic graphical models. Studies in fuzziness and soft computing, vol 213. Springer, Berlin, pp 81–102

Cooper GF, Herskovits E (1992) A Bayesian method for the induction of probabilistic networks from data. Mach Learn 9:309–347

Cowell RG, Dawid AP, Lauritzen SL, Spiegelhalter DJ (1999) Probabilistic networks and expert systems. Information science and statistics. Springer, New York

Dagum P, Luby M (1993) Approximating probabilistic inference in Bayesian belief networks is NP-hard. Artif Intell 60:141–154

Daly R, Shen Q, Aitken S (2011) Learning Bayesian networks: approaches and issues. Knowl Eng Rev 26:99–157

Darwiche A (2009) Modeling and reasoning with Bayesian networks. Cambridge University Press, Cambridge

Darwiche A (2010) Bayesian networks. Commun ACM 53:80–90

Dawid AP (2010a) Beware of the DAG! JMLR workshop Conf Proc 6:59–86

Dawid AP (2010b) Seeing and doing: the Pearlian synthesis. In: Dechter R, Geffner H, Halpern JY (eds) Heuristics, probability and causality: a tribute to Judea Pearl. College Publications, London, pp 309–325

Dethlefsen C, Højsgaard S (2005) A common platform for graphical models in R: the gRbase package. J Stat Softw 14:1–12

Didelez V, Sheehan NA (2007) Mendelian randomisation: why epidemiology needs a formal language for causality. In: Russo F, Williamson J (eds) Causality and probability in the sciences. Texts in philosophy, vol 5. College Publications, London, pp 263–292

Fast A, Hay M, Jensen D (2008) Improving accuracy of constraint-based structure learning. Technical Report 08-48, Computer Science Department, University of Massachusetts Amherst

Friedman N (1997) Learning belief networks in the presence of missing values and hidden variables. In: Fisher DH (ed) Proceedings of the fourteenth international conference on machine learning (ICML '97). Morgan Kaufmann, San Francisco, pp 125–133

Friedman N (2004) Inferring cellular networks using probabilistic graphical models. Science 303:799–805

Friedman N, Goldszmidt M, Wyner A (1999a) Data analysis with Bayesian networks: a bootstrap approach. In: Prade H, Laskey K (eds) Proceedings of the fifteenth annual conference on uncertainty in artificial intelligence (UAI-99). Morgan Kaufmann, San Francisco, pp 196–205

Friedman N, Goldszmidt M, Wyner A (1999b) On the application of the bootstrap for computing confidence measures on features of induced bayesian networks. In: Heckerman D, Whittaker J (eds) Proceedings of the seventh international workshop on artificial intelligence and statistics. Morgan Kaufmann, San Francisco, pp 197–202

Geiger D, Heckerman D, King H, Me (2001) Stratified exponential families: graphical models and model selection. Ann Stat 29:505–529

Geneletti S, Mason A, Best N (2011) Adjusting for selection effects in epidemiologic studies: why sensitivity analysis is the only "solution". Epidemiology 22:36–39

Getoor L, Rhee JT, Koller D, Small P (2004) Understanding tuberculosis epidemiology using structured statistical models. Artif Intell Med 30:233–256

Gilks WR, Richardson T, Spiegelhalter D (1996) Markov Chain Monte Carlo in practice. Chapman & Hall, Boca Raton

Glover F (1989) Tabu search – part i. ORSA J Comput 1:190–206

Glover F (1990) Tabu search – part ii. ORSA J Comput 2:4–32

Glymour C, Scheines R, Spirtes P, Ramsey J (2012) TETRAD project. http://www.phil.cmu.edu/projects/tetrad/. Accessed 15 Aug 2012

Glymour MM (2006) Using causal diagrams to understand common problems in social epidemiology. In: Oakes J, Kaufmann J (eds) Methods in social epidemiology. Jossey-Bass, San Francisco, pp 393–428

Glymour MM, Greenland S (2008) Causal diagrams. In: Rothman K, Greenland S, Lash T (eds) Modern epidemiology, 3rd edn. Lippincott Williams & Wilkins, Philadelphia, pp 183–209

Greenland S, Brumback B (2002) An overview of relations among causal modelling methods. Int J Epidemiol 31:1030–1037

Greenland S, Pearl J, Robins JM (1999) Causal diagrams for epidemiologic research. Epidemiology 10:37–48

Heckerman D (1999) A tutorial on learning with Bayesian networks. In: Jordan M (ed) Learning in graphical models. MIT, Cambridge, pp 301–354

Heckerman D, Geiger D, Chickering DM (1995) Learning Bayesian networks: the combination of knowledge and statistical data. Mach Learn 20:197–243

Hernán MA, Robins JM (2006) Instruments for causal inference: an epidemiologist's dream? Epidemiology 17:360–372

Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA (2002) Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol 155:176–184

Hernán MA, Hernández-Díaz S, Robins JM (2004) A structural approach to selection bias. Epidemiology 15:615–625

Højsgaard S (2012) Graphical independence networks with the gRain package for R. J Stat Softw 46:1–26

Højsgaard S, Edwards D, Lauritzen SL (2012) Graphical models with R. Springer, New York

Husmeier D (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics 19:2271–2282

Husmeier D (2005) Probabilistic modeling in bioinformatics and medical informatics. Springer, London

Imoto S, Goto T, Miyano S (2002) Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. Pac Symp Biocomput 7:175–186

Imoto S, Kim S, Goto T, Miyano S, Aburatani S, Tashiro K, Kuhara S (2003) Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. J Bioinform Comput Biol 1:231–252

Jensen FV, Nielsen TD (2007) Bayesian networks and decision graphs. Springer, New York

Kalisch M, Bühlmann P (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. J Mach Learn Res 8:613–636

Kalisch M, Mächler M, Colombo D, Maathuis MH, Bühlmann P (2012) Causal inference using graphical models with the R package pcalg. J Stat Softw 47:1–26

Kirkpatrick S, Gelatt CDJ, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680

Kjærulff UB, Madsen AL (2008) Bayesian networks and influence diagrams: a guide to construction and analysis. Springer, New York

Knüppel S (2011) DAG program. http://epi.dife.de/dag/. Accessed 3 Oct 2012

Knüppel S, Stang A (2010) DAG program: identifying minimal sufficient adjustment sets. Epidemiology 21:159

Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT, Cambridge

Korb KB, Nicholson AE (2011) Bayesian artificial intelligence. 2nd edn. CRC, Boca Raton

Lauritzen SL (1990) Graphical models. Clarendon, Oxford

Lauritzen SL (1992) Propagation of probabilities, means, and variances in mixed graphical association models. J Am Stat Assoc 87:1098–1108

Lauritzen SL (1995) The EM algorithm for graphical association models with missing data. Comput Stat Data An 19:191–201

Lauritzen SL, Spiegelhalter DJ (1988) Local computations with probabilities on graphical structures and their application to expert systems. J Roy Stat Soc B 50:157–224

Lauritzen SL, Dawid AP, Larsen BN, Leimer HG (1990) Independence properties of directed Markov fields. Networks 20:491–505

Li J, Wang ZJ (2009) Controlling the false discovery rate of the association/causality structure learned with the PC algorithm. J Mach Learn Res 10:475–514

Liu Z, Malone B, Yuan C (2012) Empirical evaluation of scoring functions for Bayesian network model selection. BMC Bioinform 13:S14

Lunn D, Spiegelhalter D, Thomas A, Best N (2009) The BUGS project: evolution, critique and future directions. Stat Med 28:3049–3067

Madsen AL, Lang M, , Kjærulff UB, Jensen F (2003) The Hugin tool for learning Bayesian networks. In: Nielsen TD, Zhang NL (eds) Symbolic and quantitative approaches to reasoning with uncertainty. Lecture notes in computer science, vol 2711. Springer, Berlin, pp 594–605

Markowetz F, Spang R (2007) Inferring cellular networks – a review. BMC Bioinform 8(Suppl 6):S5

Moral S, Rumí R, Salmeó A (2001) Mixtures of truncated exponentials in hybrid Bayesian networks. In: Benferhat S, Besnard P (eds) Symbolic and quantitative approaches to reasoning with uncertainty. Lecture notes in computer science, vol 2143. Springer, Berlin, pp 156–167

Murphy K (2007) Software for graphical models: a review. ISBA Bull 14:13–15

Murphy K (2012) Software packages for graphical models/ Bayesian networks. http://www.cs.ubc.ca/∼murphyk/Software/bnsoft.html. Accessed 15 Aug 2012

Nadathur SG, Warren JR (2011) Emergency department triaging of admitted stroke patients – a Bayesian network analysis. Health Inform J 17:294–312

Nguefack-Tsague G (2011) Using Bayesian networks to model hierarchical relationships in epidemiological studies. Epidemiol Health 33:e2011006

Pearl J (2009) Causality – models, reasoning and inference. 2nd edn. Cambridge University Press, Cambridge

R Development Core Team (2012) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/. Accessed 15 Aug 2012

Ramsey J (2010) Bootstrapping the PC and CPC algorithms to improve search accuracy. Tech Rep 101, Department of Philosophy, Carnegie Mellon University. http://repository.cmu.edu/philosophy/101. Accessed 15 Aug 2012

Ramsey J, Zhang J, Spirtes P (2006) Adjacency-faithfulness and conservative causal inference. In: Proceedings of the twenty-second annual conference on uncertainty in artificial intelligence (UAI-06). AUAI, Arlington, pp 401–408

Robins JM (2001) Data, design, and background knowledge in etiologic inference. Epidemiology 12:313–320

Robins JM, Blevins D, Ritter G, Wulfsohn M (1992) G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of aids patients. Epidemiology 3:319–336

Robins JM, Hernán MA, Brumback B (2000) Marginal structural models and causal inference in epidemiology. Epidemiology 11:550–560

Robins JM, Scheines R, Spirtes P, Wasserman L (2003) Uniform consistency in causal inference. Biometrika 90:491–515

Robinson R (1977) Counting unlabeled acyclic digraphs. In: Little H (ed) Combinatorial mathematics V. Lecture notes in mathematics, vol 622. Springer, Berlin, pp 28–43

Rothman KJ (1976) Causes. Am J Epidemiol 104:587–592

Rothman KJ, Greenland S, Lash T (2008) Modern epidemiology. 3rd edn. Lippincott Williams & Wilkins, Philadelphia

Rubin D (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66:688–701

Scutari M (2010) Learning Bayesian networks with the bnlearn R package. J Stat Softw 35:1–22

Shenoy PP (2011) A re-definition of mixtures of polynomials for inference in hybrid Bayesian networks. In: Liu W (ed) Symbolic and quantitative approaches to reasoning with uncertainty. Lecture notes in computer science, vol 6717. Springer, Berlin, pp 98–109

Shrier I, Platt RW (2008) Reducing bias through directed acyclic graphs. BMC Med Res Methodol 8:70

Spiegelhalter DJ, Lauritzen SL (1990) Sequential updating of conditional probabilities on directed graphical structures. Networks 20:579–605

Spirtes P, Glymour C (1990) An algorithm for fast recovery of sparse causal graphs. Report CMU-PHIL-15, Department of Philosophy, Carnegie Mellon University

Spirtes P, Meek C, Richardson T (1995) Causal inference in the presence of latent variables and selection bias. In: Besnard P, Hanks S (eds) Proceedings of the eleventh conference on uncertainty in artificial intelligence (UAI-95). Morgan Kaufmann, San Francisco, pp 499–506

Spirtes P, Glymour C, Scheines R (2001) Causation, prediction and search, 2nd edn. MIT, Cambridge

Stefanini FM, Coradini D, Biganzoli E (2009) Conditional independence relations among biological markers may improve clinical decision as in the case of triple negative breast cancers. BMC Bioinform 10(Suppl 12):S13

Textor J (2012) DAGitty v.10. http://www.dagitty.net/. Accessed 3 Oct 2012

Textor J, Hardt J, Knüppel S (2011) DAGitty: a graphical tool for analyzing causal diagrams. Epidemiology 5:745

Tsamardinos I, Brown LE, Aliferis CF (2006) The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn 65:31–78

VanderWeele TJ, Robins JM (2007a) Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. Am J Epidemiol 166:1096–1104

VanderWeele TJ, Robins JM (2007b) Four types of effect modification: a classification based on directed acyclic graphs. Epidemiology 18:561–568

Verma T, Pearl J (1991) Equivalence and synthesis of causal models. In: Bonissone P, Henrion M, Kanal L, Lemmer J (eds) Proceedings of the sixth conference on uncertainty in artificial intelligence (UAI-90). Elsevier, Amsterdam, pp 258–268

Verma T, Pearl J (1992) An algorithm for deciding if a set of observed independencies has a causal explanation. In: Dubois D, Wellman MP, D'Ambrosio B, Smets P (eds) Proceedings of the eighth conference on uncertainty in artificial intelligence (UAI-92). Morgan Kaufmann, San Mateo, pp 323–330

Wang M, Chen Z, Cloutier S (2007) A hybrid Bayesian network learning method for constructing gene networks. Comput Biol Chem 31:361–372

Weinberg CR (1993) Toward a clearer definition of confounding. Am J Epidemiol 137:1–8

Weinberg CR (2007) Can DAGs clarify effect modification? Epidemiology 18:569–572

Wong ML, Lee SY, Leung KS (2002) A hybrid approach to discover Bayesian networks from databases using evolutionary programming. In: Proceedings of the 2002 IEEE international conference on data mining, ICDM ’02. IEEE Computer Society, Los Alamitos, pp 498–505

Part IV

Exposure-Oriented Epidemiology

Life Course Epidemiology

40

Yoav Ben-Shlomo, Gita Mishra, and Diana Kuh

Contents

Y. Ben-Shlomo (✉)
School of Social and Community Medicine, University of Bristol, Bristol, UK

G. Mishra
School of Population Health, University of Queensland, Brisbane, Australia

D. Kuh
MRC Unit for Lifelong Health and Ageing, University College London, London, UK

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_56,

40.1 Introduction and Brief History

Life course epidemiology is the study of long-term biological, behavioral, and psychosocial processes that link adult health and disease risk to physical or social exposures acting during gestation, childhood, adolescence, and earlier or adult life or across generations (Kuh and Ben-Shlomo 2004). Life course epidemiology was one of several new conceptual models of epidemiological thinking that began to emerge in the 1980s and 1990s and that are now mainstream paradigms in social epidemiology (Susser 1985; Susser and Susser 1996a; Krieger and Zierler 1997; McMichael 1999), though its concepts have been applied more generally to chronic disease etiology. They parallel the emergence of causal models (Greenland et al. 1999; Hernán et al. 2002) for epidemiology during the same period.

Life course epidemiology started in the United Kingdom, when a group of epidemiologists and social scientists[1] started to meet regularly at the London School of Hygiene and Tropical Medicine to evaluate the new empirical evidence on the role of early life influences, particularly fetal development, on chronic disease risk that was emerging from the Medical Research Council (MRC) Environmental Epidemiology Unit (Southampton) under the leadership of Prof. David Barker. This body of work, initially known as the "Barker" hypothesis, fundamentally asserted that the causes of many, if not all, chronic diseases were established by "programming" of processes during embryological development and this in turn was influenced by nutritional influences acting either in pregnancy or prepregnancy through the development of the mother during her childhood or adolescence (Barker 1998). The programming hypothesis itself evolved into a more generalized developmental model that extended into the postnatal and childhood period (Gluckman and Hanson 2004). Initial reactions within epidemiology to these new ideas were either skeptical or at best agnostic, although their challenge to existing conventional life style models of chronic disease was welcomed. We (YBS and DK), with our colleagues, seized the opportunity to provide a critical overview of these new ideas and review of the research findings. This led to the first edition of the Oxford University Press publication "A life course approach to chronic disease epidemiology" (Kuh and Ben-Shlomo 1997) which is currently in its second edition (Kuh and Ben-Shlomo 2004) and has led to a series of other life course books on women's health (Kuh and Hardy 2003), methodology (Pickles et al. 2007), family-based designs (Lawlor and Mishra 2009a) and healthy ageing (Kuh et al. 2014). We coined the term "life course epidemiology" to reduce the polarization that was occurring between epidemiologists supporting either the early-life or adult-life hypotheses on the development of chronic disease risk,

[1] The group included Mel Bartley, Yoav Ben-Shlomo, David Blane, Derek Cook, George Davey Smith, Jonathan Elford, Diana Kuh, Dave Leon, Ivan Perry, Chris Power, David Strachan, and Peter Whincup

arguing that the relative importance of exposures across life needed to be investigated and underlying processes examined.

In the course of putting together the first book and a related editorial (Ben-Shlomo and Kuh 2002) and glossary (Kuh et al. 2003), two of us (YBS, DK) were forced to specify a clear conceptual framework behind some of our rather loose terminology and causal thinking. We proposed several potential models or pathways through which early or later life exposures may work in either an additive or interactive fashion to increase individual and population risk of disease. These models, based on ideas of critical and sensitive periods or of risk accumulation from other disciplines (see Sect. 40.2), were our first attempt to establish a conceptual foundation for others to extend and/or reformulate. As such, they provided a useful stepping stone for hypothesis formulation.

This chapter allows us to develop these models as well as acknowledge and synthesize the ideas of other researchers. We place life course epidemiology within an historical and interdisciplinary perspective. We restate our initial models as well as define and clarify key terminology used in life course epidemiology and often a source of confusion between epidemiologists and other disciplines. Discussion of the uses and misuses of life course models in the published literature then helps us to refine our initial models. We discuss both design and analytical issues for those who wish to undertake life course research and devote a section around the less commonly used but long-established family-based designs explaining how they help infer causality. We illustrate some of our ideas with two case studies around life course influences on an age-related disease and trait.

We have encountered three common criticisms to life course epidemiology. The first is “Haven’t we always been doing this?” As we shall describe below, the broad ideas behind life course epidemiology are not new and have been considered in some guise or other for at least a century both within public health and other social sciences. However, the theoretical specifications and empirical data supporting these ideas are indeed contemporary. The second criticism is “Isn’t it obvious that all disease must stem from exposures acting across the life course?” This is also a truism for any multifactorial chronic disease, but as we have previously noted, “it is insufficient to glibly state that all health and social outcomes are due to life course influences. This is analogous to stating that all health is a function of genetic and environmental exposures. Whilst factually correct it does not further our understanding of etiology or help policy formulation” (Ben-Shlomo and Kuh 2004). We feel that this asks the wrong question. Instead, as is elucidated in our models, the key questions relate to whether certain time periods are more relevant for some exposures, whether exposure duration is the main determinant, and how exposures may be interrelated as part of causal pathways. The third criticism is “These models are too complex. Is it realistic, either in terms of data availability or statistical models to truly test them?” This last criticism is the most challenging. We hope that the examples provided at least partially answer this criticism and will enthuse other epidemiologists to build on our ideas and state their life course hypotheses in a more thoughtful manner.

40.2 Life Course Epidemiology Within Its Wider Historical and Inter-Disciplinary Perspectives

Epidemiology and public health have historically gone through various eras or phases of development from the nineteenth century until today (Hamlin 1992; Susser and Susser 1996b), each emphasizing a different notion of health, a different notion of cause or set of causative factors, different research settings, and different intervention strategies to prevent or reduce risks to health. We have suggested (Kuh and Davey Smith 1993, 2004; Kuh et al. 2009) that in one of these eras, at the beginning of the twentieth century, health was seen as fitness or vitality, and there was concern that the fitness of the population was degenerating. The contemporary debate centered around whether fitness was inherited or could be nurtured by a good early environment, especially by "maternal efficiency" (Pearson 1919; Paton and Findlay 1926). Public health officials who believed that "the health of the adult is dependent upon the health of the child" and that "the health of the child is dependent upon the health of the infant and its mother" (Newman 1914, p. 16) argued that the new maternal and infant welfare services would improve national fitness; the geneticists argued that these services would keep "weaklings" alive and lead to reduced national fitness. There was little epidemiological evidence for either side. Rather support for the importance of the early environment came from the biological and psychological sciences where ideas of developmental critical periods were gaining ground (Stockard 1927; Freud 1955). Later, it came from early cohort analysis that showed birth cohort effects on mortality risk (Kermack et al. 1934), although the importance of these has been recently challenged (Murphy 2010).

In epidemiology and public health, interest in early life influences of adult health waned as early post war cohort studies of middle aged men successfully identified more proximal biological risk factors, such as high blood pressure, and adult lifestyle factors, for adult chronic disease. In contrast, early life continued to dominate the psychological sciences into the 1970s and then gave way to a life-span perspective. Some emphasized the role of continuing adult development on later life outcomes (Baltes et al. 1998), and others showed discontinuity as well as continuity between childhood and adult psychopathology and demonstrated the presence of "chains of risk" and how they could be broken (Rutter 1989; Rutter et al. 2006). The modern revival of the life course perspective in human biology and biological anthropology linked early development to aging and emphasized how the early as well as the later environment influences individual variation in structure and function at every age. The notion of risk accumulation, while present in early public health and demographic literature (Ciocco et al. 1941; Riley 1989), has become more prevalent with the transcendence of the disposable soma theory of biological aging where aging is seen to be caused by the accumulation of molecular and cellular damage from insults experienced by the organism throughout life (Kirkwood and Holliday 1979).

Epidemiologists were relatively late in reviving a life course perspective. The catalyst was the growing empirical evidence of a multiplicity of links between the postnatal as well as the prenatal environment and developmental characteristics with respect to adult health, function, disease, and aging. Evidence came from population studies, initially from the maturing birth cohort studies (Wadsworth 1991; Power and Elliott 2006) and the revitalized historical cohort studies (Barker 1992) and more recently from intergenerational studies (see Sect. 40.5). Systematic reviews and meta-analyses have also strengthened the evidence base (Huxley et al. 2000, 2002).

In recent years, this epidemiological evidence has been harnessed in support of the biological processes of developmental plasticity. The fundamental assumption is that multiple phenotypes can arise during development from a single genotype, which enables production of a range of phenotypes better suited to the prevailing environmental conditions than would otherwise be possible (Gluckman et al. 2007, 2008) but may lead to disease or disadvantage if environmental boundaries are exceeded. Epigenetic mechanisms are thought to mediate these processes (Gluckman et al. 2009). Developmental plasticity is seen as an evolutionary process that helps to explain how organisms become matched to face environmental challenges. This occurs on several time scales: across multiple generations (natural selection), across the life course (developmental plasticity), and across shorter time periods (homeostasis).

Epidemiologists thus have an increasingly rich set of theories of human development and aging on which to draw to explain as well as describe the phenomena they observe. These theories should underpin the building of conceptual epidemiological models that hypothesize the relationships between risk factors across life and health outcomes in later life. Epidemiologists must then find study designs to operationalize these relationships into testable hypotheses and, ideally, to distinguish causal relationships. If empirical evidence supports the hypothesis, scientists must then explain the phenomena in ways that are compatible with or further develop these underlying theories. Research hypotheses and findings are often not refined sufficiently to distinguish between competing theories, and a number of interpretations may be plausible. Or they represent only pieces of the jigsaw that can be reinterpreted or more fully interpreted when other jigsaw pieces are available (see Sect. 40.6). As many studies simply do not have data across the whole of life, findings from different studies may need to be pieced together using advanced statistical modeling (Wills et al. 2011). This can be facilitated by standardized and repeated measures of biological function (such as blood pressure, muscle strength, lung function, or cognitive function) that exhibit age-related change across the life course and are thus dynamic tools for investigating links between development and aging across life and the biological imprint of physical and social exposures (Kuh 2007).

Life course epidemiologists who wish to explain how a distal risk factor, such as father's social class, in childhood may affect lifetime biological function and the development of adult chronic disease risk can draw on several theoretical disciplines: sociological or economic theories to explain the social and economic structure of society, psychological theories to explain individual agency and behavior, and biological theories to explain how social and psychological phenomena "get under the skin." More often than not, they focus on one pathway of interest (the "direct" pathway) and ignore the others.

40.3 Life Course Models

“All models are wrong; the practical question is how wrong do they have to be to not be useful” (cited from Box and Draper (1987) in Collins and Altman (2010)).

40.3.1 Basic Concepts

The pathways which link an exposure such as poor socioeconomic conditions in childhood to a pathological entity such as a cerebrovascular accident many years later are complex and may involve factors acting at both societal, area, neighborhood, family, individual, and ultimately physiological and cellular levels (Susser and Susser 1996a). While the models we propose below are gross simplifications of these complex processes, they support the empirical test of alternative or counterfactual models of disease risk.

We initially proposed a simple dichotomous classification of life course models which conceptualized the origins of chronic disease to belong to either a critical period or accumulation of risk pathway (we shall criticize this classification below). Within these two subdivisions, we further elucidated various alternative models to be considered. We illustrated the latter set of models using simple graphical illustrations (more formally referred to as directed acyclic graphs – DAGs) but did not do this for the first set as we were uncertain as to the best way to illustrate these graphically (the problem of illustrating interactions in DAGs has only been recently addressed) (Vanderweele et al. 2010). We use hypothetical or empirically based examples to illustrate each of these models.

1. *Critical Period Model* (see Box 40.1 for definitions of key terms). In this model, it is assumed that an exposure, e.g. a specific viral infection, acting within a predefined time window, e.g. first trimester of pregnancy, may result in a specific disorder, e.g. schizophrenia. This model does not preclude other causes of schizophrenia that are unrelated to viral infection, e.g. genetic factors, or that other exposures, e.g. expressed familial hostility, cannot also increase risk of schizophrenia. It does assume that maternal exposure after the first trimester or if a subject is exposed postnatally is not associated with any increased risk compared to an unexposed subject.
2. *Sensitive Period Model*. An obvious variant of the above scenario is that exposure in the second trimester is associated with a milder form of the disease and that exposure in the third trimester is associated with a schizophreniform type of personality but not overt clinical disease. This is what is meant by a sensitive rather than a critical period model, namely, that exposure within specific time windows has a greater or lesser risk of disease. Both these models can be regarded as examples of age interaction whereby the effect of viral infection is modified by the time window (as measured by gestational age) of exposure. In a critical period model, the effect is all or none by time strata, while in the sensitive period model, it is simply that the strength of the relationship is stronger in one time period than another.

Box 40.1. Glossary of key concepts

Accumulation of risk – The notion that life course exposures or insults gradually accumulate through episodes of illness and injury, adverse environmental conditions, and health-damaging behaviors.

Chain of risk model – A chain of risk model refers to a sequence of linked exposures that raise disease risk because one bad experience or exposure tends to lead to another and then another. Social, biological, and psychological chains of risk are possible and involve "mediating factors" and often "modifying factors." The sequential links are probabilistic rather than deterministic.

Critical period – In the natural sciences, a critical period of development refers to a time window when intrinsic changes in the organization of living systems or subsystems toward increasing complexity, greater adaptivity, and more efficient functioning occur rapidly and may be most easily modified in a favorable or unfavorable direction. In life course epidemiology, a critical period is a limited time window in which an exposure can have adverse or protective effects on development and subsequent disease outcome. Outside this developmental window, there is no excess disease risk associated with exposure.

Interaction – Interaction or effect modification is the process whereby a third variable ("effect modifier") alters the relationship between an exposure and an outcome so that the effect of an exposure may only be seen, for example, in those who are also exposed to the effect modifier. It is known as "synergism" if the modifying variable enhances the effect of the exposure or "antagonism" if it diminishes it.

Sensitive period – Like critical periods, sensitive periods are also times of rapid individual change, but there is more scope to modify or even reverse those changes outside the time window. Thus, a sensitive period in life course epidemiology is a time period when an exposure has a stronger effect on development and subsequent disease risk than it would at other times. Outside the time period, any excess risk will be weaker.

Adapted from Kuh et al. (2003)

3. *Critical/Sensitive Period Model with Later Effect Modification*. Even though the timing of an exposure may be essential to have any effect or a stronger effect on a disease outcome, this does not mean that it is inevitable that disease will always emerge. Unrelated exposures in later life that are not themselves secondary to the initial insult could still modify disease risk either through independent or interactive effects. For example, if maternal undernutrition in pregnancy, as seen in the Dutch Hunger Winter (Stein et al. 1975), was associated with a permanent reduction in the number of muscle cells at birth, this could be compensated by

increased physical activity in later life resulting in muscle hypertrophy. In this case, reduced grip strength would only be observed in small babies who were relatively inactive as adults or with aging, and there may be no effect in those with high activity levels.

4. *Accumulation of Risk with Uncorrelated Exposures.* We live in an associational world whereby many non-genetic exposures are correlated to varying degrees due to social patterning of exposures (Davey Smith et al. 2007). It is therefore rare to have environmental associations that are truly randomly distributed. In such a scenario, the risk of being exposed to A is unrelated to B and C so that unexposed individuals are equally as likely to be exposed to B and C (see Fig. 40.1a). For example, the risk of depression may be associated with a genetic variant (A), death of a father due to military conflict, and unemployment in adult life (C) due to the subject's employer going bankrupt. In this example, there is no reason to believe any of these exposures are correlated with each other. If each exposure increases risk (though this may be to varying degrees), then individuals exposed to more than one factor will have a greater risk than those exposed to fewer factors. In this case, we are assuming that these exposures do not interact with each other and the effects are additive.
5. *Accumulation of Risk with Correlated Exposures.* Exposures are more commonly correlated because of risk clustering (see Fig. 40.1b). For example, living in a poor neighborhood may be associated with being exposed to a less healthy diet, reduced opportunities to exercise, and greater peer influences on smoking. Each of these factors may additively increase risk of coronary heart disease, but in this case, one exposure will be associated with the others due to the common factor of neighborhood poverty which is an upstream determinant of the other mediating factors.
6. *Chain of Risk Additive Model.* In this scenario, each exposure increases risk, but A is itself a determinant of B which in turn increases the risk of C (see Fig. 40.1c). Hence, a chain effect may be established whereby an exposure may only have a modest effect directly on outcome, but its overall effect, including the indirect pathways, may be much larger. For example, smoking may directly result in subclinical atherosclerosis through an inflammatory effect on the arterial wall, but it also reduces exercise behavior due to respiratory symptoms. This in turn results in reduced aerobic capacity but also increases obesity. This in turn results in insulin resistance syndrome and is a risk factor for coronary heart disease. In this case, intervening on obesity alone would have health benefits, but obese individuals who have followed this pathway would still be at increased risk due to their life course history compared to non-obese individuals.
7. *Chain of Risk Trigger Model.* In contrast to the previous model, the trigger model is only associated with risk due to the last exposure in the chain (see Fig. 40.1d). For example, poor childhood socioeconomic environment results in overcrowding which in turn increases exposure to *Helicobacter pylori*, and this is associated with peptic ulcer. In this scenario, eradication of *H. pylori* will eliminate the risk, and there will no longer be any residual risk associated with the socioeconomic environment. Such pathways are interesting as they may

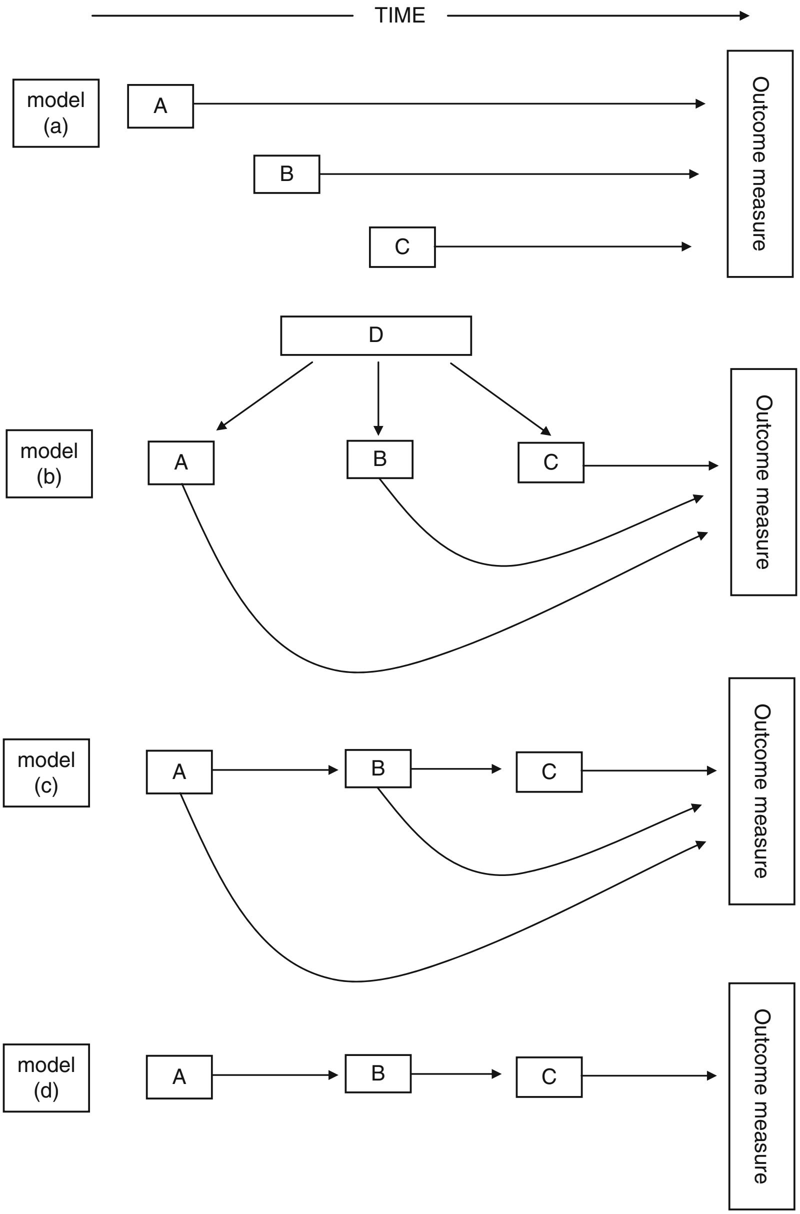


Fig. 40.1 Graphical illustrations of various life course models: (**a**) accumulation of risk with uncorrelated exposures, (**b**) accumulation of risk with correlated exposures, (**c**) chain of risk additive model, and (**d**) chain of risk trigger model (Taken from Kuh et al. 2003)

explain why some epidemiological studies show marked heterogeneity of effects (assuming this is not due to random variation). In these pathways, variations in temporal or cultural differences can abolish associations. So, for example, the association of socioeconomic conditions with any smoking-mediated disease can legitimately vary by time and place due to cultural or temporal changes in smoking behavior.

We did not invent the concepts of accumulation and chains of risk but rather adopted them from existing ideas. The demographer John Riley proposed the idea of accumulation as one of the driving forces behind patterns of mortality, so that countries with greater adverse exposures exposed individuals to an accumulation of adversity and this resulted in higher mortality (Riley 1989). Similarly, the eminent child psychiatrist Sir Michael Rutter coined the term chains of risk so that "the impact of some factor in childhood may lie less in the immediate behavioral change it brings about than in the fact it sets into motion a chain reaction in which one 'bad' thing leads to another, or, conversely, that a good experience makes it more likely that another one will be encountered." (Rutter 1989).

40.3.2 Misunderstandings Behind Life Course Epidemiology

The ideas behind life course epidemiology have been embraced to varying degrees across a wide range of topic areas from cardiovascular, cancer, mental health, and aging. Common misunderstandings frequently occur around careless use of the terminology, for example, the use of the term "critical period" when in fact the term sensitive period may be more appropriate. For example, it has been argued that the elasticity of the arterial vessels is determined in both pre- and postnatal life (Martyn and Greenwald 1997), and so an adverse exposure acting in this period could result in stiffer arteries which in turn result in adult hypertension. In such a scenario, poor prenatal growth (a proxy measure of an adverse exposure) may be more sensitive than poor postnatal growth, but the intrauterine period cannot be critical if effects of adverse development and growth can still be seen due to postnatal exposures. If one expanded the biological time window to include both pre- and postnatal life, assuming no influences outside this time window, then by definition this would be a critical period. However, this is unhelpful, and the a priori definition of what time windows one should select should be based on either biological evidence and/or differences in the determinants of pre- and postnatal growth and hence the ability to potentially intervene.

Similarly, the demonstration of a stronger association in an earlier period conditional on the same exposure in a later time period does not necessarily support a "programming effect" in the earlier period. This issue was initially highlighted in relation to the inverse association between birth weight and later cardiovascular outcomes in some cases only after conditioning on a measure of later adiposity. We shall discuss this example further in relation to modeling trajectories.

40.3.3 Social Mobility and Life Course Models

When we conceptualized our models, we had assumed that the critical/sensitive period models applied to a single exposure that could occur across different time windows. In this case, the question is whether the timing of exposure has additional effects over and above the duration of exposure. In contrast, our accumulation models were based on different exposures acting either in an additive or chain of risk fashion. However, the accumulation model has most often been applied to the same exposure, e.g. socioeconomic position (SEP), over time so that the risk of an outcome has been compared by simply adding the number of periods of poor SEP and examining if risk increases in a simple linear fashion (Davey Smith et al. 1997). Given that earlier SEP is a strong predictor of future SEP, this would better fit an additive chain of risk model than the other types of model. However, sociologists have had a long interest in the process of social mobility so it was not surprising that a publication attempted to address whether the effects of critical period, accumulation, and social mobility in terms of CHD risk could be disentangled (Hallqvist et al. 2004). Hallqvist and colleagues examined SEP (which was dichotomized into manual vs. non-manual) over three time periods (childhood, young adulthood, and midlife). They then identified eight possible patterns which were classified as consistent with either a critical period, accumulation, or social mobility model. However, given that the social mobility trajectory was consistent with both the critical period and accumulation trajectories, they concluded that it was impossible to disentangle these different models analogous to the problem of trying to disentangle age period and cohort effects. We support this conclusion but feel that this problem occurred because they added social mobility as an additional model. As can be seen in Table 40.1, the phenomenon of social mobility is consistent with a critical/sensitive period or accumulation model depending on which of many empirically based patterns emerge from the data analysis. We would argue that social mobility is better limited to a descriptive term regarding inter- and intra-general variation in SEP as opposed to an etiological model which tries to understand how the process of mobility is embodied into pathophysiology. A more recent approach has systematically tried to test whether one can differentiate between these different models having explicitly operationalized each model in

Table 40.1 The association between social mobility and disease risk operating under different life course models

Early SEP	Late SEP	Effect estimates				
High	High	1.0	1.0	1.0	1.0	1.0
Low	High	1.5	2.0	1.0	1.8	1.5
High	Low	1.5	1.0	2.0	1.3	1.0
Low	Low	2.0	2.0	2.0	2.0	5.0
Life course model		Accumulation	Early critical period	Late critical period	Early sensitive period	Interaction model

algebraic terms (Mishra et al. 2009). This later paper also used three measures of SEP over time but importantly defined social mobility in a specific manner so that it reflected an interactive effect of mobility, thus resolving the intractable problem seen in the earlier paper. In this paper, the authors included an adult-life critical period model, though it is less likely from a biological perspective that exposures in adult life would be critical as in general terms this reflects a degenerative rather than a developmental phase of the life span.

This approach has been an important advance in helping researchers formally test different models where exposures are categorized. Other approaches are required for trajectories of continuous measures (see below). One simple problem that has already been identified is that the timing of the measurement points could artefactually favor one model over another. In a recent publication examining SEP and physical performance measures, only two time points were available: childhood and adult life. Simple multivariable analyses suggested that the effect of adult SEP was a stronger determinant of poor performance than childhood SEP supporting an adult-life sensitive period model rather than a simple accumulation chain of risk model (Birnie et al. 2010). Formal tests however did not reveal a significantly better fit between these two models probably due to insufficient power. However, the adult SEP exposure was based on a measure that covered a much longer time period than the childhood measure. This was not an issue in the original paper by Mishra and colleagues, as it used measures that spanned three relatively equal-spaced time periods. In the latter example, the greater effect of adult SEP may not reflect any greater sensitivity but merely a better measure of exposure as it captures socioeconomic conditions that operate over a longer time period.

A similar problem has also been encountered when trying to tease out sensitive and accumulation models in relation to the duration of effect. A publication from the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort examined whether pre- and postnatal maternal anxiety predicted asthma risk. It found that children whose mothers reported anxiety in both periods had a greater odds ratio of asthma than those only reporting anxiety in either period alone (odds ratio for both periods 1.46, 95% confidence interval (CI) = 1.20–1.78; prenatal only odds ratio 1.30, 95% CI = 1.04–1.64; postnatal only odds ratio 1.18, 95% CI = 0.89–1.57) (Cookson et al. 2009). At first glance, these results appear to be most compatible with an accumulation effect: however, further exploration of this relationship suggested that duration of anxiety symptoms was actually a marker of prenatal severity. Women reporting both pre- and postnatal anxiety symptoms had higher anxiety scores at 32 weeks gestation than those with just prenatal anxiety. The duration of symptoms across both time periods was thus a measure of prenatal severity, and conditioning on prenatal anxiety abolished the association with postnatal anxiety, more consistent with a critical or sensitive period model.

Even without some of the issues highlighted above, there is a danger that applying these statistical approaches without any understanding of the biological basis of disease may still produce misleading results. A paper examining smoking habits in women in relation to breast cancer suggested that initiation of smoking in puberty may be related to a greater risk of breast cancer after taking smoking duration

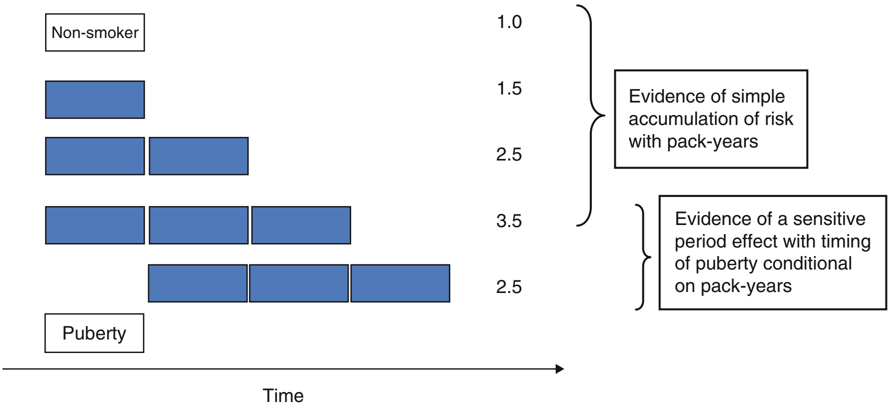

Non-smoker
1.0
1.5
2.5
3.5
2.5
Evidence of simple accumulation of risk with pack-years
Evidence of a sensitive period effect with timing of puberty conditional on pack-years
Puberty
Time

Fig. 40.2 The association between cigarette pack-years and breast cancer risk (each *dark block* indicates 10 pack-years)

into account (Band et al. 2002). In the illustrated hypothetical example, we have partitioned exposure time into 10 pack-year blocks (Fig. 40.2). A naive analysis would conclude that a simple accumulation model was the best fit of the data given the steady increase (dose response effect) in risk of breast cancer. However, by comparing the subgroup of women with 30 pack-years who started smoking around the time of puberty with those who started after puberty, we would now conclude that puberty may be a sensitive period where, conditional on the total pack-years of exposure, women may have an additional greater risk possibly due to the enhanced carcinogenic effects of tobacco exposure on rapidly dividing breast tissue.

It is clear that if one thinks of accumulation in terms of the same rather than different exposures, then the critical and sensitive period models are not qualitatively different but could be considered to be embedded (Fig. 40.3) within a hierarchical framework. Thus, a sensitive model is a variation of the accumulation model whereby exposures accumulate risk over time, but the effects are differential and the critical period model is a special form of a sensitive period model whereby you have an all or none effect. It may therefore be more helpful to consider an accumulation model as the default model with sensitive and critical period models considered as special types accumulation models.

40.3.4 The Importance of Trajectories

We have described above how patterns of a dichotomous variable over time may have differential effects on disease risk. The same applies to continuous measures such as anthropometry. The initial birth weight chronic disease/risk factor associations which stimulated such huge interest in biological programming were mainly based on only examining intrauterine growth through the proxy variable of

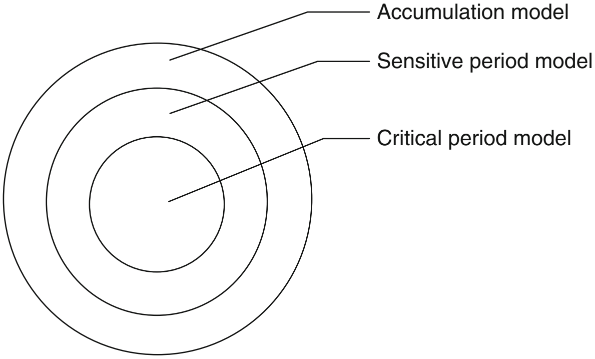


Fig. 40.3 A hierarchical representation of life course models

birth weight, length, or ponderal index (though the original Hertfordshire cohort did also examine anthropometry at 1 year). However, growth does not stop at birth and may be considered a life course process until maturity. Because fetal growth is so fundamental to embryological development, it is seductive to only consider this period as “critical” which may be true for some outcomes, e.g. thalidomide and limb development, but not for others, e.g. lung function, arterial elasticity, and synaptic density where postnatal influences are also important. The interpretation of the original inverse association between birth weight and later life outcomes as an effect of intrauterine development was criticized as in many (but not all) publications: this association only became clear after statistical adjustment for a later life measure of anthropometry such as weight or body mass index. This conditional association suggested that postnatal centile crossing or catch-up growth may be the important pathway rather than intrauterine development, which has rather different policy implications if true (Lucas et al. 1999). The biological interpretation of this statistical finding is more complex. It is possible that centile crossing is itself driven by poor intrauterine growth. This may be through an acute phenomenon, e.g. infants whose mothers smoked in pregnancy may develop suboptimally but once born catch up to their genetic potential. Alternatively endocrine and/or metabolic pathways may be reset having a longer term chronic effect on growth trajectories. Another alternative explanation is that both pre- and postnatal growth may be determined by genetic factors which in turn may be related to later disease (Hattersley and Tooke 1999). This example highlights the importance of understanding trajectories over time rather than examining one or two arbitrary time points that are available in individual datasets.

The analysis of repeat measures to model trajectories is complex and open to various different analytical strategies. Initial publications from the retrospective Helsinki cohort used a simple summary measure approach so that standardized differences in anthropometry, measured on an annual basis between 6 and 16 years, were compared between CHD cases and the whole cohort (Eriksson et al. 1999). Such patterns are useful descriptive starting points but may be misleading as they imply an individual trajectory from what are repeat cross-sectional measures. Repeat longitudinal measures of the same individuals can be used to derive summary measures. For example, a study examining growth between 9 and 18 years and with up to 65 repeat anthropometric measures was able to construct individual growth velocity curves and hence derive from the inflection point of maximal growth velocity, the age at peak height velocity, which was used as a valid surrogate for the timing of puberty (Sandhu et al. 2006). A more sophisticated reanalysis of the same data using a multilevel model approach was able to explain 99% of the variance in growth patterns. This SuperImposition by Translation And Rotation (SITAR) model (Cole et al. 2010) derived three parameters (starting point size – equivalent to the intercept, growth tempo – a left-right shift in the spline curve and growth velocity – a scaling parameter that reflects the duration of the growth spurt making the curves steeper or shallower) that were able to capture how individuals deviated from the overall mean growth curve. This method ideally requires repeat data closely spaced together. Multivariable models have also been used to model growth velocities over more widely spaced time periods which have the advantage of being less correlated than the crude anthropometric measures measured over shorter periods (Kuh et al. 2006); these can show very high correlation coefficients. Other approaches have used latent variable models or spline regression models. In each case, these models are testing whether deviations from a "normative" trajectory are associated with an outcome of interest. The spline approach has one useful advantage in that it allows periods of time to be partitioned, a priori, based on biologically driven changes in growth patterns (Ben-Shlomo et al. 2008). For example, postnatal growth is non-linear, and there are transitions or inflection points of deceleration or acceleration. Having modeled such points, the effect of each period independent of the earlier period can be investigated. The use of directed acyclic graphs can also highlight the direct and indirect influence of each period, as normal regression models will only present the direct estimates conditional on other parameters.

The ability to test various trajectories is however limited by the degree of variability within any dataset. A recent study examining the associations between the age at onset of obesity and later blood pressure found evidence, particularly in men, that earlier onset of obesity was associated with both higher body mass index at 53 years and higher blood pressure (Wills et al. 2010). However, given the strong tracking of body mass index (BMI) (among individuals who first became obese at 26 years, 82%, 91%, and 95% were obese at 36 years, 43 years, and 53 years), they could not test patterns that included weight loss subsequent to obesity in earlier life as there were few individuals in these trajectories. While one could argue that this may not matter if few individuals demonstrate these trajectories, the counterfactual still exists, and if an effective intervention was to be implemented (assuming it has equal benefits at all time periods), then we would still remain unclear as to its optimum timing.

We have briefly mentioned some of the statistical challenges to be considered within a life course framework (for a fuller discussion, see Pickles et al. 2007). A review of the statistical issues around life course analyses concluded that there was no one “correct” approach to conduct a specific analysis but that the method should be driven by the specific question being tested and the available data (De Stavola et al. 2006). However, life course analyses have the same problems as any longitudinal analysis, that is loss to follow-up bias and missing data, and these too need careful attention.

40.4 General Design Issues

Ideally one should have access to a multigenerational birth cohort to test life course hypotheses. While in the United Kingdom, we are fortunate to have several birth cohorts (NSHD, NCDS, ALSPAC, millennium cohort), and it is possible to link family members across generations in several European countries, none contain the ideal dataset. For example, the oldest birth cohort studies may not have biological samples from earlier periods, and more recent birth cohorts are still relatively young and hence only have intermediate surrogates of disease risk. There is therefore no perfect and complete life course study, and even if there was, its generalizability could be limited because of the effects of social change. Therefore, life course hypotheses can be tested, and valuable insights gained using historical cohorts, case control studies, and rarely but more convincingly through the follow-up of natural experiments (Lumey et al. 1993; Stanner et al. 1997) or randomized controlled trials, usually set up for other short-term outcomes. In addition, combining trajectories from several cohorts using multilevel models is also starting to reveal valuable insights on the lifetime trajectories of biological functions such as blood pressure which are important drivers of disease risk (Wills et al. 2011).

We highlight a few examples of these approaches for illustrative purposes. The role of perinatal risk factors has played a prominent role in ideas around programming. However, exposures such as maternal diet and smoking during pregnancy will be heavily confounded by other socially patterned variables. One approach to this problem has been to compare maternal-offspring to paternal-offspring associations (see Sect. 40.5 below for more details) on the basis that intrauterine effects should be specific to the former while effects due to confounding will find associations with both parents (Brion et al. 2007). However, an alternative strategy is to break the confounding using a natural experiment. In a developing world context, this can be explored by examining season of birth as this relates to conception and fetal development during dry or rainy seasons that have a large effect on maternal nutrition. A subtle variant of this approach has examined the effects of day-time fasting during Ramadan on the birth weights of offspring and other abnormalities. This is particularly clever as the Muslim calendar varies by season so that in some years, the duration will be shorter or longer depending on whether it falls during winter or summer. Thus, mothers who fasted during the first trimester were more likely to have lighter birth weights, a reduction of male offspring, and a greater risk of disability (Almond and Mazumder 2008). An extreme version of this approach can be seen in studies that have followed up offspring of mothers who conceived during a period of starvation, e.g. Dutch Hunger Winter, Leningrad siege, and Chinese famine. Any nutritional hypothesis is severely challenged if such studies fail to find effects, even though one can question the generalizability of these experiences.

Another approach has been the longer term follow-up of past randomized trials. Such an approach is dependent on several key factors: (a) does the intervention relate to an a priori hypothesis, (b) did it actually have an effect, (c) can the original participants be traced, and (d) is follow-up sufficiently complete and unbiased to enable valid comparisons, ideally as an intention to treat analysis, by random group allocation. For example, Singhal and colleagues have developed the "growth acceleration hypothesis" (Singhal and Lucas 2004) on the basis of their follow-up studies of premature babies who were randomized to either breast versus formula feeding or high versus low nutrient feeds. In these studies, the response rate at follow-up was low (around 20%), and though there do not appear to be major differences at baseline between those who reattended compared to those who did not at baseline, there remains a concern that confounding may still be present though to a lesser degree. However, many of their reported associations are using the data as a cohort rather than using the initial randomization thereby losing the benefits of the trial design. One of us (YBS) has been involved in the follow-up of village-based cluster trial in India where a nutritional intervention was introduced to exposed villages while this was delayed for several years in the control villages (Kinra et al. 2008). Follow-up of the children from all the villages has enabled a pragmatic analysis of whether intervention was associated with better or worse cardiovascular risk markers. Reassuringly, children from the intervention villages were taller at follow-up, indicating some effect on development, and had better insulin function as well as less stiff arteries, though the differences were modest. Whether these observations translate into more meaningful differences in future risk of hypertension and diabetes remains to be seen, and further more detailed phenotypic follow-up has just been completed.

40.5 Family-Based Designs

Families lie at the center of life course epidemiology, but their role is only beginning to be reflected sufficiently in our understanding of individuals not as isolated beings but as lives that unfold within collective structures that shape our health over and above individual characteristics (Merlo et al. 2009). As has been discussed in previous sections, maternal influences during gestation can play a key role in human development, but mothers alongside other family members also characterize the experience of childhood and adolescence. Once in adulthood, partners, pregnancies, and offspring will condition an individual's health and health-related behaviors. The effects on health and well-being due to relationships with adult offspring and grandchildren may also extend into later life. Thus, family-based studies may be seen as representing another opportunity for life course epidemiology in moving beyond associational epidemiology to multilevel analyses that address questions of causality in the absence of randomized control trials, disentangling genetic and environmental factors, and the role of timing of exposures.

The theory and practice of family-based studies are set out in a recent publication from the life course series entitled "Family matters: Designing analysing and understanding family-based studies in life course epidemiology" (Lawlor and Mishra 2009a). This examines the three key aspects of how family-based analyses can contribute to life course epidemiology. "First, family will directly affect health by determining many of the biological, environmental and social exposures across the life course. Secondly, different family members will exert their impact on ones health by different degrees at different times in the life course, and hence detailed studies of family influences could help to understand the importance of timing of exposures across the life course. Lastly, comparing statistical relationships within and between families can help to clarify the mechanisms underlying associations in life course studies and help to determine causality" (Lawlor and Mishra 2009b).

Family-based studies are those designs that incorporate purposeful data collection on family members. They include the long-established design of twin studies, which have provided first estimates for effects of heredity compared with environmental influences, as well as sibling studies where the effects of different exposures can be examined for individuals who are genetically very similar but not identical. Over recent years, there has been an increasing focus on the use of large population-based studies that cover multiple generations, for instance, by providing prospective data on parents and offspring from registry records. Thus, these studies present opportunities for investigating causality by allowing us to examine intra-familial correlation of outcomes, such as for birth weight over generations, while accounting for shared environments.

As discussed in the previous sections regarding the developmental origins hypothesis, a key aspect is the possibility of maternal exposures during pregnancy having a direct biological effect of offspring outcomes. Findings, however, can be heavily confounded by social and environmental factors. One approach to this issue using intergenerational studies has been to compare maternal-offspring to paternal-offspring associations for the effect of smoking on offspring birth weight and childhood BMI (Davey Smith 2008). Epidemiological findings have established that smoking during pregnancy by the mother, but not paternal smoking, leads to lower birth weight and thus provide evidence for an intrauterine mechanism (Lawlor and Mishra 2009b). Maternal smoking during pregnancy is also linked with childhood obesity (Oken et al. 2008), but elsewhere higher (rather than lower) birth weight has been associated with childhood obesity (Baird et al. 2005), implying that a distinct mechanism exists for the effect of smoking on birth weight. Moreover, from the Avon Longitudinal Study of Parents and Children, the effect of maternal smoking on childhood obesity has been shown as similar in magnitude to paternal smoking (Leary et al. 2006). This suggests that maternal smoking does not have a direct intrauterine effect on childhood obesity but may be due to the confounding effects of other lifestyle factors (such as diet and exercise) of parents who smoke. Further research is needed, but study sample size is an issue in having sufficient numbers of families with either or both parents smoking, including those where mothers (temporarily or permanently) gave up smoking during pregnancy.

Even with family-based studies, causality remains difficult to nail down. In a German longitudinal study of substance use and mental disorders, findings confirmed previous results that major depression among parents was associated with increased offspring risk for depression (Lieb et al. 2002). While such results support the potential role of genetic and various shared environmental factors leading to mental disorder, the study was unable to eliminate the possibility for the association to operate in the other direction, for instance, psychopathology in offspring leading to depression in their parents. The analysis to support specific mechanisms for depression is further complicated since mental disorders often do not occur singularly but as a group of comorbid disorders.

Family-based studies pose a number of methodological challenges beyond those of representativeness normally encountered in life course epidemiology, such as missing data and attrition. Considerable caution needs to be applied in generalizing results from long-term birth cohort studies, since cohort effects may limit the extent that older individuals in the sample are representative of the contemporary population, especially of younger generations. But in addition, we need to consider the various ways that families – and the social environment they provide – may have changed over decades, including the increased prevalence of single parent families, the rise of working mothers, and even the decline in the number of siblings. Assumptions that siblings have the same fixed familial socioeconomic position may break down in countries, cultures, and populations where one child is “favored” over others, such as in their diet or for further education, based on their gender, birth order, early signs of promise, or other characteristics (Mishra and Lawlor 2009). It is also important to consider the validity and potential bias of using proxy informants in family-based studies, whereby the respondent gives information about other family members, such as when a mother provides data about her offspring or vice versa. Moreover, when family structure and/or interrelationships are central to a particular analysis, missing data from one family member may mean that available data from other family members should not be used.

Sample size and statistical power are often an issue in the design of family-based studies. When they are used to test causal hypotheses or disentangle causal pathways, they typically involve comparing variations in the associations within and between different family structures, for example, comparing maternal-offspring associations to paternal-offspring associations or comparing within and between siblings or twins. Such analyses require considerably greater sample sizes than studies concerned with a single association, and they are prone to type 2 errors (Mishra et al. 2009). In cases where family-based studies are analyzed as Mendelian randomization studies (Davey Smith and Ebrahim 2003) (where genetic variants that are known to be robustly associated with a non-genetic risk factor of interest are treated as an instrumental variable for that risk factor), the maternal genotype is often used as an instrumental variable for an intrauterine exposure. These studies require very large sample sizes, typically tens of thousands of families depending on the prevalence of the genetic variant under investigation, in order to have the statistical power to provide estimates that would be useful in translating research into public health policy or clinical practice.

The increasing focus on family-based studies in life course epidemiology has seen existing cohort studies augmented with data collection on family members or through record linkage of registry data as well as medical records (for instance, cancer registry, DNA data, and health-care utilization). The designs of some new cohort studies address family associations from the outset, such as the Generation Scotland study which will recruit adults over a 6-year period, to form a total of 50,000 individuals that comprise of siblings and parent-offspring groups (Smith et al. 2006). In another example, the *China Children and Families Cohort Study* has enrolled 300,000 families with recruitment occurring even prior to conception of the main participant for some couples (Brown et al. 2007). The LifeLines project aims to recruit 165,000 individuals in the Netherlands over three generations including partners and siblings (Stolk et al. 2008). Thus, the future for family-based studies lies both in the prospect of new family data from existing studies and the commencement of studies with younger cohorts – especially those now being initiated in low- and middle-income countries. If studies are conducted mindful of the diligence and methodological rigor gained from existing cohort research, then over time this should allow for cross cohort comparisons to test and verify key epidemiological findings and disease mechanisms across generations and populations with diverse cultural, socioeconomic, and environmental conditions.

40.6 Life Course and Aging: Two Case Studies

This section will use two examples relevant to understanding lifetime influences of aging to highlight some of the interesting questions raised by a life course approach. Aging toward the end of life is commonly associated with the phenomenology of “frailty” defined as a syndrome with some or all of the following manifestations: weakness, fatigue, weight loss, decreased balance, low levels of physical activity, slowed motor processing and performance, social withdrawal, mild cognitive changes, and increased vulnerability to stressors. A multiplicity of etiological factors have been suggested for frailty (Hogan et al. 2003), including the potential role of the growth hormone (GH) and insulin-like growth factor (IGF) axis. IGF is a nutritional-driven peptide that is an insulin analogue and has pleiotropic actions including positive effects on mitogenesis, cytodifferentiation, and anti-apoptosis. It is the last role that has generated much interest in relation to cancer etiology, with a meta-analysis highlighting its potential role in premenopausal breast cancer and possibly prostate cancer (Renehan et al. 2004).

The GH-IGF axis also plays an important role in both fetal and postnatal growth, and animal studies show marked reductions in genetic mutation models. For example, mutations in the IGF-I and IGF-II genes can reduce birth weight by 35–40% and adult body size by 60–70% (Veldhuis et al. 2006). Therefore, this axis may be an important link between development and aging, in that more rapid growth and early maturation will be associated with shorter life expectancy given prominent evolutionary drive for reproduction. We have previously shown that earlier puberty is associated with higher IGF-I levels in adulthood measured 40 years later (Sandhu et al. 2006), using a historical cohort design. However, in this study, we had no measure of IGF during childhood so we could not differentiate as to whether higher childhood IGF-I triggered earlier puberty and then tracked into adulthood or whether the timing of puberty itself had an effect on the GH-IGF axis. The modifiability of the system as well as its potential long-term patterning however was seen in the follow-up of the Barry Caerphilly milk supplementation trial (Elwood et al. 1981). This study randomized women in pregnancy and their offspring to additional free household milk and measured growth up to 5 years. While intervention was associated with higher birth weight, there was no demonstrable effect on childhood height at 5. However, follow-up at 25 years found that adults in the intervention group had lower IGF-I levels suggesting that supplementation may have downregulated the GH-IGF axis (Ben-Shlomo et al. 2005). This complemented a report from the Dutch Hunger Winter where women who were exposed to nutritional restriction as children had higher IGF-I levels in middle age, suggesting an upregulation of the GH-IGF axis (Elias et al. 2004). A similar finding was seen in the Boyd Orr cohort where subjects whose family reported greater consumption of milk had lower IGF-I levels in adulthood (Martin et al. 2007). These findings were unexpected as one would have predicted, a priori, the opposite results: acute supplementation or restriction is associated with an increase or decrease of IGF-I levels, respectively. However, acute changes may reverse over the longer term as evidenced by a small follow-up of children who had IGF measured in the first year of life and then again at 17 years (Larnkjaer et al. 2009). This study observed an inverse association between postnatal levels and IGF levels at 17 years. This observation may explain the earlier paradoxical findings. An alternative explanation is that subjects with higher IGF-I simply mature faster and have already reached the declining phase of IGF-I levels that is associated with deceleration of the pubertal growth spurt so that they appear to have lower IGF-I levels in early adulthood in comparison to those who are maturing more slowly and reach their peak IGF-I level at a later age. This observation may no longer be true in much later life.

IGF-I levels may provide one possible mechanism that explains a series of observations relating to breast cancer and measures of physical performance (one constituent of the frailty syndrome). A meta-analysis of birth weight and breast cancer found a continuous positive association (dos Santos Silva et al. 2008), and a large historical cohort study found that independent of birth, taller stature at 8 and 14 years and more rapid pubertal growth increased risk of breast cancer (Ahlgren et al. 2004). Similarly, data from the MRC National Survey of Health and Development has found that childhood weight gain before age 7 years, in contrast to adult weight gain, was generally beneficial for midlife physical performance (timed tests of balancing on one leg or rising from a chair and sitting down ten times), though this finding was seen in men and not in women (Kuh et al. 2006). This finding is consistent with a meta-analysis of studies that show that better childhood SEP is associated with better performance measures over and above adult SEP (Birnie et al. 2011) and lower IGF-I is associated with lower SEP both at birth and in adulthood

(Kumari et al. 2008). Data from the Caerphilly Prospective Study (CaPS) has also shown that higher IGF-I predicts better physical performance (as measured by the get up and go test) almost 2 decades later (Birnie – personal communication).

This example shows how these findings from different studies are like pieces of a jigsaw puzzle that, if put together with the help of an underlying conceptual and theoretical model, can reveal some of the potential pathways linking development and aging. Given the complexity of most endocrine systems and their ability to change due to negative feedback and up/downregulation, it is perhaps not unsurprising that there may be early life effects on the GH-IGF axis with short-term implications for development and long-term implications for aging and age-related diseases.

Our second example also demonstrates the need to consider developmental trajectories. In almost all societies, it is recognized that systolic blood pressure increases with age. This effect is even more marked if one examines pulse pressure (difference between systolic and diastolic pressure) or pulse wave velocity (McEniery et al. 2010a), a more direct measure of arterial stiffness. This is unsurprising as the wall of arteries consists of smooth muscle, collagen, and elastin tissue, and every time the heart beats, the arterial wall expands and recoils which puts stress on its constituent structures. A recent 20-year follow-up from CaPS found that the best predictor of arterial stiffness was the combined product of heart rate and systolic blood pressure measured over 25 years, and most of the usual cardiovascular risk factors had only modest effects if at all (McEniery et al. 2010b). In this example, one might regard the increase of stiffness with aging as a simple accumulation effect so that the more cardiac cycles one experiences at greater pressure, the more likely there will be physical damage to the arteries and hence greater stiffness, though this is still modifiable to some degree due to greater physical fitness and pharmacological lowering of blood pressure and heart rate. However, it is also clear that the development of the arterial tree, which is both a pre- and postnatal phenomenon, could contribute to future stiffness by altering the ratio of collagen to elastin tissue (Martyn and Greenwald 1997). One way to examine this idea is to look at another physiological measure that declines with age and also develops during the same time period. Tests of lung function, as a measure of the pulmonary tree, may be one such measure, and these are inversely correlated with arterial stiffness so that subjects with better lung function have less stiff arteries. What is more surprising is that a past measure of lung function in CaPS was a better predictor of pulse wave velocity than a cross-sectional measure taken at the same time (Bolton et al. 2009) and that rate of decline in lung function had little predictive effect. This suggests that there are common developmental origins of lung function and arterial structure that could influence both and that these may be more important than smoking history on arterial stiffness. In this example, both sensitive and accumulation models probably fit the data, but they operate through different mechanisms. These findings illustrate the importance of thinking about both risk factors that determine the development of a function as well as those that may determine its decline. In this hypothetical figure (Fig. 40.4), we have illustrated four population groups: a normative group who live up to a 100 years without

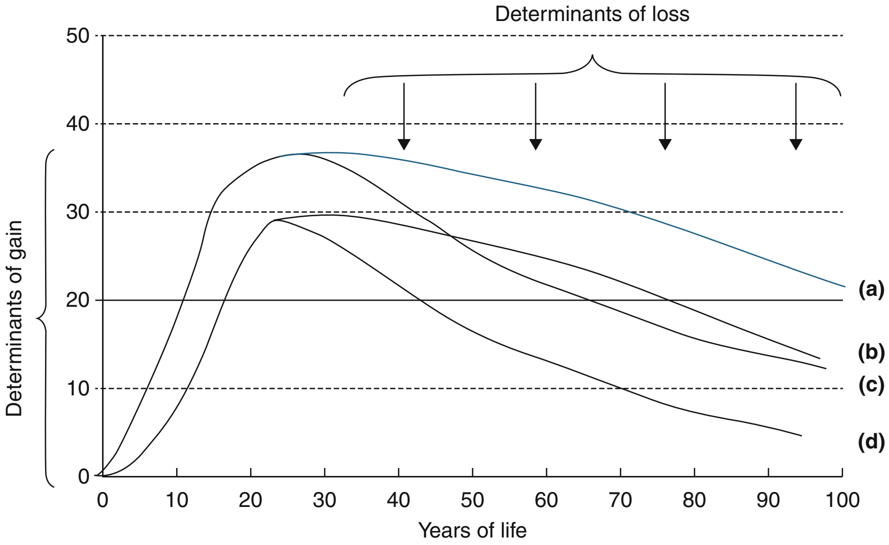


Fig. 40.4 Population trajectories of traits across the life course illustrating four potential groups: (*a*) normative, (*b*) suboptimal development, (*c*) accelerated decline, and (*d*) combined developmental and decline group

reaching the threshold for disease (a), a suboptimal developmental group (b), an accelerated decline group (c), and a combined group (d). By only studying disease or a phenotype at the end of these various pathways, one will attenuate associations. Other than the combined group which will have an earlier age at onset, the pathways (b) and (c) will be clinically indistinguishable unless one has longitudinal data over the life course that can identify which pathway has been followed. Thus, a risk factor that acts on suboptimal development (groups b and d) will appear to have an attenuated effect when combining groups (b and c) depending on their respective frequencies in the sample. The longitudinal approach is therefore more powerful to elucidate different classes of risk factors.

40.7 Conclusions and Policy Implications

Life course epidemiology is predicated on the idea that what happens to individuals early in life can have long-term effects for later health, disease, and aging. Understanding the lifetime influences on health, disease, and aging will require an interdisciplinary life course approach which is being fostered across the UK research councils. For too long, the various life course perspectives across the disciplines have traveled parallel but separate paths. There have been growing calls for life-span psychology and life course sociology to integrate ideas of social structure and individual agency (Settersten 2009), but more attention by these disciplines needs to be paid to relevant biological theories. Biologists and epidemiologists have generally tried to control out the social and psychological pathways and in the UK at least have focused on specific diseases, rather than underlying traits or disease-related trajectories.

Life course ideas have stimulated not only academics but policy makers, and various NGOs (such as the World Health Organization and specific disease charities) (Aboderin et al. 2002; National Heart Forum 2003) have embraced these concepts. This has surprised us as many questions remain unanswered especially in relation to cost-effectiveness of interventions. In addition, most life course interventions will only demonstrate benefits over a long time frame, usually of little interest to politicians. However, they do allow policy makers some flexibility as interventions at any point across the life course can be argued to have long, medium, or short-term benefits. In particular, the degree to which inequalities in health can be reduced by life course models has received some attention. A commentary in the Journal of the American Medical Association argued that "from both basic research and policy perspectives, confronting the origins of disparities in physical and mental health early in life may produce greater effects than attempting to modify health-related behaviors or improve access to health care in adulthood" (Shonkoff et al. 2009). The Marmot Review of Health Inequalities *Fair Society, Healthy Lives* (Marmot Review Team 2010) highlighted the importance of interventions in early life but took a more long-term accumulation perspective. "Central to the Review is a life course perspective. Disadvantage starts before birth and accumulates throughout life,... Meanwhile, there is much that can be done to improve the lives and health of people who have already reached school, working age and beyond .... Services that promote the health, well being and independence of older people and, in so doing, prevent or delay the need for more intensive or institutional care, make a significant contribution to ameliorating health inequalities." As we have argued above, both approaches may be correct in some respects. Given our current state of knowledge in relation to human intervention studies, we are operating in a "speculation rich, data sparse" environment. The beginning of the twenty-first century offers many challenges and exciting opportunities to develop life course perspectives for a greater understanding of the etiology of disease and health and for the rationale development of public health prevention programs.

References

Aboderin I, Kalache A, Ben-Shlomo Y, Lynch JW, Yajnik CS, Kuh D, Yach D (2002) Life course perspectives on coronary heart disease, stroke and diabetes: key issues and implications for policy and research. World Health Organization, Geneva

Ahlgren M, Melbye M, Wohlfahrt J, Sorensen TI (2004) Growth patterns and the risk of breast cancer in women. N Engl J Med 351:1619–1626

Almond D, Mazumder B (2008) Health capital and the prenatal environment: the effect of maternal fasting during pregnancy. National Bureau of Economic Research, Boston, NBER Working Paper No. W14428

Baird J, Fisher D, Lucas P, Kleijnen J, Roberts H, Law C (2005) Being big or growing fast: systematic review of size and growth in infancy and later obesity, BMJ 331:929

Baltes PB, Lindenberger U, Staudinger UM (1998) Life-span theory in developmental psychology. In: Damon W, Lerner RM (eds) Handbook of child psychology volume 1: theoretical models of human development. Wiley, New York, pp 1029–1143

Band PR, Le ND, Fang R, Deschamps M (2002) Carcinogenic and endocrine disrupting effects of cigarette smoke and risk of breast cancer. Lancet 360:1044–1049

Barker DJP (1992) Fetal and infant origins of adult disease. BMJ Publishing Group, London

Barker DJP (1998) Mothers, babies and health in later life, 2nd edn. Churchill Livingstone, London

Ben-Shlomo Y, Kuh D (2002) A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. Int J Epidemiol 31:285–293

Ben-Shlomo Y, Kuh D (2004) Conclusions. In: Kuh D, Ben-Shlomo Y (eds) A life course approach to chronic disease epidemiology, 2nd edn. Oxford University Press, Oxford, pp 446–463

Ben-Shlomo Y, Holly J, McCarthy A, Savage P, Davies D, Davey Smith G (2005) Prenatal and postnatal milk supplementation and adult insulin-like growth factor I: long-term follow-up of a randomized controlled trial. Cancer Epidemiol Biomark Prev 14:1336–1339

Ben-Shlomo Y, McCarthy A, Hughes R, Tilling K, Davies D, Davey Smith G (2008) Immediate postnatal growth is associated with blood pressure in young adulthood: the Barry Caerphilly Growth Study. Hypertension 52:638–644

Birnie K, Cooper R, Martin RM, Kuh D, Sayer AA, Alvarado BE, Bayer A, Christensen K, Cho SI, Cooper C, Corley J, Craig L, Deary IJ, Demakakos P, Ebrahim S, Gallacher J, Gow AJ, Gunnell D, Haas S, Hemmingsson T, Inskip H, Jang SN, Noronha K, Osler M, Palloni A, Rasmussen F, Santos-Eggimann B, Spagnoli J, Starr J, Steptoe A, Syddall H, Tynelius P, Weir D, Whalley LJ, Zunzunegui MV, Ben-Shlomo Y, Hardy R, HALCyon study team (2011) Childhood socioeconomic position and objectively measured physical capability levels in adulthood: a systematic review and meta-analysis. PLoS One 6:e15564

Birnie K, Martin RM, Gallacher J, Bayer A, Gunnell D, Ebrahim S, Ben-Shlomo Y (2010) Socio-economic disadvantage from childhood to adulthood and locomotor function in old age: a lifecourse analysis of the Boyd Orr and Caerphilly prospective studies. J Epidemiol Community Health. doi: 10.1136/jech.2009

Bolton CE, Cockcroft JR, Sabit R, Munnery M, McEniery CM, Wilkinson IB, Ebrahim S, Gallacher JE, Shale DJ, Ben-Shlomo Y (2009) Lung function in mid-life compared with later life is a stronger predictor of arterial stiffness in men: the Caerphilly Prospective Study. Int J Epidemiol 38:867–876

Brion MJ, Leary SD, Davey Smith G, Ness AR (2007) Similar associations of parental prenatal smoking suggest child blood pressure is not influenced by intrauterine effects. Hypertension 49:1422–1428

Brown RC, Dwyer T, Kasten C, Krotoski D, Li Z, Linet MS, Olsen J, Scheidt P, Winn DM (2007) Cohort profile: the International Childhood Cancer Cohort Consortium (I4C). Int J Epidemiol 36:724–730

Ciocco A, Klein H, Palmer CE (1941) Child health and the selective service physical standards. Public Health Rep 56:2365–2375

Cole TJ, Donaldson MD, Ben-Shlomo Y (2010) SITAR–a useful instrument for growth curve analysis. Int J Epidemiol 39:1558–1566

Collins GS, Altman DG (2010) An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. BMJ 340:c2442

Cookson H, Granell R, Joinson C, Ben-Shlomo Y, Henderson AJ (2009) Mothers' anxiety during pregnancy is associated with asthma in their children. J Allergy Clin Immunol 123:847–853

Davey Smith G (2008) Assessing intrauterine influences on offspring health outcomes: can epidemiological findings yield robust results? Basic Clin Pharmacol Toxicol 102:245–256

Davey Smith G, Ebrahim S (2003) Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol 32:1–22

Davey Smith G, Hart C, Blane D, Gillis C, Hawthorne V (1997) Lifetime socioeconomic position and mortality: prospective observational study. BMJ 314:547–552

Davey Smith G, Lawlor DA, Harbord R, Timpson N, Day I, Ebrahim S (2007) Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. Plos Med 4:e352

De Stavola BL, Nitsch D, dos Santos Silva I, McCormack V, Hardy R, Mann V, Cole TJ, Morton S, Leon DA (2006) Statistical issues in life course epidemiology. Am J Epidemiol 163:84–96

dos Santos Silva I, De SB, McCormack V (2008) Birth size and breast cancer risk: re-analysis of individual participant data from 32 studies. Plos Med 5:e193

Elias SG, Keinan-Boker L, Peeters PH, Van Gils CH, Kaaks R, Grobbee DE, van Noord PA (2004) Long term consequences of the 1944–1945 Dutch famine on the insulin-like growth factor axis. Int J Cancer 108:628–630

Elwood PC, Haley TJL, Hughes SJ, Sweetnam PM, Gray OP, Davies DP (1981) Child growth (0–5 years), and the effect of entitlement to a milk supplement. Arch Dis Child 56:831–835

Eriksson JG, Forsén T, Tuomilehto J, Winter PD, Osmond C, Barker DJP (1999) Catch-up growth in childhood and death from coronary heart disease: longitudinal study. BMJ 318:427–431

Freud S (1955) The claims of psycho-analysis to scientific interest: the interest of psycho-analysis from a developmental point of view (1913). In: Strachey J (ed) The complete psychological works of Sigmund Freud, vol XIII. Hogarth Press, London, pp 182–184

Gluckman PD, Hanson MA (2004) Living with the past: evolution, development, and patterns of disease. Science 305:1733–1736

Gluckman PD, Hanson MA, Beedle AS (2007) Early life events and their consequences for later disease: a life history and evolutionary perspective. Am J Hum Biol 19:1–19

Gluckman PD, Hanson MA, Beedle AS, Spencer HG (2008) Predictive adaptive responses in perspective. Trend Endocrinol Metab 19:109–110

Gluckman PD, Hanson MA, Buklijas T, Low FM, Beedle AS (2009) Epigenetic mechanisms that underpin metabolic and cardiovascular diseases. Nat Rev Endocrinol 5:401–408

Greenland S, Pearl J, Robins JM (1999) Causal diagrams for epidemiologic research. Epidemiology 10:37–48

Hallqvist J, Lynch J, Bartley M, Lang T, Blane D (2004) Can we distangle life course processes of accumulation, critical period and social mobility? An analysis of disadvantaged socioeconomic positions and myocardial infarction in the Stockholm Heart Epidemiology Program. Soc Sci Med 58:1555–1562

Hamlin C (1992) Predisposing causes and public health in early nineteenth-century medical thought. Soc Hist Med 5:43–70

Hattersley A, Tooke JE (1999) The fetal insulin hypothesis: an alternative explanation of the association of low birth weight with diabetes and vascular disease. Lancet 353:1789–1792

Hernán MA, Hernández-Diaz S, Werler MM, Mitchell AA (2002) Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol 155:176–184

Hogan DB, MacKnight C, Bergman H (2003) Models, definitions, and criteria of frailty. Aging Clin Exp Res 15:1–29

Huxley R, Neil A, Collins R (2002) Unravelling the "fetal origins" hypothesis: is there really an inverse association between birth weight and future blood pressure? Lancet 360:659–665

Huxley RR, Shiell AW, Law CM (2000) The role of size at birth and postnatal catch-up growth in determining systolic blood pressure: a systematic review of the literature. J Hypertens 18: 815–831

Kermack WO, McKendrick AG, McKinlay PL (1934) Death-rates in Great Britain and Sweden. Some general regularities and their significance. Lancet 1:698–703

Kinra S, Rameshwar Sarma KV, Ghafoorunissa, Mendu VV, Ravikumar R, Mohan V, Wilkinson IB, Cockcroft JR, Davey Smith G, Ben-Shlomo Y (2008) Effect of integration of supplemental nutrition with public health programmes in pregnancy and early childhood on cardiovascular risk in rural Indian adolescents: long term follow-up of Hyderabad nutrition trial. BMJ 337:a605

Kirkwood TB, Holliday R (1979) The evolution of ageing and longevity. Proc R Soc Lond B Biol Sci 205:531–546

Krieger N, Zierler S (1997) The need for epidemiologic theory. Epidemiology 8:212–214

Kuh D (2007) A life course approach to healthy aging, frailty, and capability. J Gerontol A Biol Sci Med Sci 62:717–721

Kuh D, Ben-Shlomo Y (1997) A life course approach to chronic disease epidemiology. Oxford University Press, Oxford

Kuh D, Ben-Shlomo Y (2004) A life course approach to chronic disease epidemiology, 2nd edn. Oxford University Press, Oxford

Kuh D, Davey Smith G (1993) When is mortality risk determined? Historical insights into a current debate. Soc History Med 6:101–123

Kuh D, Davey Smith G (2004) The life course and adult chronic disease: an historical perspective with particular reference to coronary heart disease. In: Kuh D, Ben-Shlomo Y (eds) A life course approach to chronic disease epidemiology, 2nd edn. Oxford University Press, Oxford, pp 15–40

Kuh D, Hardy R (2003) A life course approach to women's health. Oxford University Press, Oxford

Kuh D, Ben-Shlomo Y, Lynch J, Hallqvist J, Power C (2003) Life course epidemiology. J Epidemiol Community Health 57:778–783

Kuh D, Hardy R, Butterworth S, Okell L, Richards M, Wadsworth M, Cooper C, Sayer AA (2006) Developmental origins of midlife physical performance: evidence from a British birth cohort. Am J Epidemiol 164:110–121

Kuh D, Hardy R, Hotopf M, Lawlor DA, Maughan B, Westendorp RJ, Cooper R, Black S, Mishra G (2009) A review of lifetime risk factors for mortality. Br J Actuar 15(Supplement):17–64

Kuh D, Cooper R, Hardy R, Richards M, Ben-Shlomo Y (eds) (2014) A life course approach to healthy ageing. Oxford University Press, Oxford

Kumari M, Tabassum F, Clark C, Strachan D, Stansfeld S, Power C (2008) Social differences in insulin-like growth factor-1: findings from a British birth cohort. Ann Epidemiol 18:664–670

Larnkjaer A, Ingstrup HK, Schack-Nielsen L, Hoppe C, Molgaard C, Skovgaard IM, Juul A, Michaelsen KF (2009) Early programming of the IGF-I axis: negative association between IGF-I in infancy and late adolescence in a 17-year longitudinal follow-up study of healthy subjects. Growth Horm IGF Res 19:82–86

Lawlor DA, Leary S, Davey Smith G (2009) Theoretical underpinning for the use of intergenerational studies in life course epidemiology. In: Lawlor DA, Mishra GD (eds) Family matters: designing, analysing and understanding family-based studies in life course epidemiology. Oxford University Press, Oxford, pp 11–38

Lawlor DA, Mishra GD (2009a) Family matters: designing, analysing and understanding family-based studies in life course epidemiology. Oxford University Press, Oxford

Lawlor DA, Mishra GD (2009b) Why family matters: introduction. In: Lawlor DA, Mishra GD (eds) Family matters: designing, analysing and understanding family-based studies in life course epidemiology. Oxford University Press, Oxford, pp 1–7

Leary SD, Davey Smith G, Rogers IS, Reilly JJ, Wells JC, Ness AR (2006) Smoking during pregnancy and offspring fat and lean mass in childhood. Obesity (Silver.Spring) 14:2284–2293

Lieb R, Isensee B, Hofler M, Pfister H, Wittchen HU (2002) Parental major depression and the risk of depression and other mental disorders in offspring: a prospective-longitudinal community study. Arch Gen Psychiatry 59:365–374

Lucas A, Fewtrell MS, Cole TJ (1999) Fetal origins of adult disease-the hypothesis revisited. BMJ 319:245–249

Lumey LH, Ravelli ACJ, Wiessing LG, Koppe JG, Treffers PE, Stein ZA (1993) The Dutch famine birth cohort study: design, validation of exposure, and selected characteristics of subjects after 43 years follow-up. Paediatr Perinat Epidemiol 7:354–367

Marmot Review Team (2010) Fair society, health lives: the Marmot Review – strategic review of health inequalities in England post-2010. The Marmot Review, London

Martin RM, Holly JM, Middleton N, Davey Smith G, Gunnell D (2007) Childhood diet and insulin-like growth factors in adulthood: 65-year follow-up of the Boyd Orr Cohort. Eur J Clin Nutr 61:1281–1292

Martyn CN, Greenwald SE (1997) Impaired synthesis of elastin in walls of aorta and large conduit arteries during early development as an initiating event in pathogenesis of systemic hypertension. Lancet 350:953–955

McEniery CM, Yasmin, Maki-Petaja KM, McDonnell BJ, Munnery M, Hickson SS, Franklin SS, Cockcroft JR, Wilkinson IB, Anglo-Cardiff Collaboration Trial Investigators (2010a) The impact of cardiovascular risk factors on aortic stiffness and wave reflections depends on age: the Anglo-Cardiff Collaborative Trial (ACCT III). Hypertension 56:591–597

McEniery CM, Spratt M, Munnery M, Yarnell J, Lowe GD, Rumley A, Gallacher J, Ben-Shlomo Y, Cockcroft JR, Wilkinson IB (2010b) An analysis of prospective risk factors for aortic stiffness in men: 20-year follow-up from the Caerphilly prospective study. Hypertension 56:36–43

McMichael AJ (1999) Prisoners of the proximate: loosening the constraints on epidemiology in an age of change. Am J Epidemiol 149:887–897

Merlo J, Ohlsson H, Lynch KF, Chaix B, Subramanian SV (2009) Individual and collective bodies: using measures of variance and association in contextual epidemiology. J Epidemiol Community Health 63:1043–1048

Mishra GD, Lawlor DA (2009) The future of family-based studies in life course epidemiology: challenges and opportunities. In: Lawlor DA, Mishra GD (eds) Family matters: designing, analysing and understanding family-based studies in life course epidemiology. Oxford University Press, Oxford, pp 325–333

Mishra G, Nitsch D, Black S, De SB, Kuh D, Hardy R (2009) A structured approach to modelling the effects of binary exposure variables over the life course. Int J Epidemiol 38:528–537

Murphy M (2010) Reexamining the dominance of birth cohorts effects on mortality. Popul Develop Rev 36:365–390

National Heart Forum (2003) Lifecourse approach to coronary heart disease prevention: a scientific and policy review. Stationery Office Books, London

Newman G (1914) Annual report for 1913 of chief medical officer of the board of education, Cmnd 7330. His Majesty’s Stationery Office, London

Oken E, Levitan EB, Gillman MW (2008) Maternal smoking during pregnancy and child overweight: systematic review and meta-analysis. Int J Obes (Lond) 32:201–210

Paton DN, Findlay L (1926) Child life investigations. Poverty, nutrition and growth. Studies of child life in cities and rural districts of Scotland. HMSO, London

Pearson K (1919) The intensity of natural selection in man. Proc Roy Soc Lond 85B:469–476

Pickles A, Maughan B, Wadsworth M (2007) Epidemiological methods in life course research, 1st edn. Oxford University Press, Oxford

Power C, Elliott J (2006) Cohort profile: 1958 British birth cohort (National Child Development Study). Int J Epidemiol 35:34–41

Renehan AG, Zwahlen M, Minder C, O’Dwyer ST, Shalet SM, Egger M (2004) Insulin-like growth factor (IGF)-1, IGF binding protein-3, and cancer risk: systematic review and meta-regression analysis. Lancet 363:1346–1353

Riley JC (1989) Sickness, recovery and death: a history and forecast of ill health, vol 1. MacMillan, London, pp 1–295

Rutter M (1989) Pathways from childhood to adult life. J Child Psychol Psychiatry 30:23–51

Rutter M, Kim-Cohen J, Maughan B (2006) Continuities and discontinuities in psychopathology between childhood and adult life. J Child Psychol Psychiatry 47:276–295

Sandhu J, Davey Smith G, Holly J, Cole TJ, Ben-Shlomo Y (2006) Timing of puberty determines serum insulin-like growth factor-I in late adulthood. J Clin Endocrinol Metab 91:3150–3157

Settersten RA Jr (2009) It takes two to tango: the (un)easy dance between life-course sociology and life-span psychology. Adv Life Course Res 14:74–81

Shonkoff JP, Boyce WT, McEwen BS (2009) Neuroscience, molecular biology, and the childhood roots of health disparities: building a new framework for health promotion and disease prevention. J Am Med Assoc 301:2252–2259

Singhal A, Lucas A (2004) Early origins of cardiovascular disease: is there a unifying hypothesis? Lancet 363:1642–1645

Smith BH, Campbell H, Blackwood D, Connell J, Connor M, Deary IJ, Dominiczak AF, Fitzpatrick B, Ford I, Jackson C, Haddow G, Kerr S, Lindsay R, McGilchrist M, Morton R, Murray G, Palmer CN, Pell JP, Ralston SH, St Clair D, Sullivan F, Watt G, Wolf R, Wright A, Porteous D, Morris AD (2006) Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. BMC Med Genet 7:74

Stanner SA, Bulmer K, Andrès C, Lantseva OE, Borodina V, Poteen VV, Yudkin JS (1997) Does malnutrition in utero determine diabetes and coronary heart disease in adulthood? Results from the Leningrad siege study, a cross sectional study. BMJ 315:1342–1349

Stein Z, Susser M, Saenger G, Marolla F (1975) Famine and human development. The Dutch Hunger Winter of 1944–1945. Oxford University Press, London

Stockard CR (1927) Hormones and structural development. The Beaumont Foundation Lecture Series 6. Wayne County Medical Society, Detroit

Stolk RP, Rosmalen JG, Postma DS, de Boer RA, Navis G, Slaets JP, Ormel J, Wolffenbuttel BH (2008) Universal risk factors for multifactorial diseases: LifeLines: a three-generation population-based study. Eur J Epidemiol 23:67–74

Susser M (1985) Epidemiology in the United States after world war II: the evolution of technique. Epidemiol Rev 7:147–177

Susser M, Susser E (1996a) Choosing a future for epidemiology: II. From black box to Chinese boxes and eco-epidemiology. Am J Pub Health 86:674–677

Susser M, Susser E (1996b) Choosing a future for epidemiology: I. Eras and paradigms. Am J Pub Health 86:668–673

Vanderweele TJ, Vansteelandt S, Robins JM (2010) Marginal structural models for sufficient cause interactions. Am J Epidemiol 171:506–514

Veldhuis JD, Roemmich JN, Richmond EJ, Bowers CY (2006) Somatotropic and gonadotropic axes linkages in infancy, childhood, and the puberty-adult transition. Endocr Rev 27:101–140

Wadsworth MEJ (1991) The imprint of time: childhood, history and adult life. Oxford University Press, Oxford

Wills AK, Hardy RJ, Black S, Kuh DJ (2010) Trajectories of overweight and body mass index in adulthood and blood pressure at age 53: the 1946 British birth cohort study. J Hypertens 28:679–686

Wills AK, Lawlor DA, Matthews F, Ahie Sayer A, Bakra E, Ben-Shlomo Y, Benzeval M, Brunner E, Cooper R, Kivimaki M, Kuh D, Muniz-Terrera G, Hardy R (2011) Life course trajectories of systolic blood pressure using longitudinal data from eight UK cohorts. PLoS Med 8(6):e1000440 (http://dx.doi.org/10.1371/journal.pmed.1000440)

Social Epidemiology 41

Tarani Chandola, Meena Kumari, and Michael Marmot

Contents

T. Chandola (✉)
CCSR and Social Statistics, University of Manchester, Manchester, UK

M. Kumari • M. Marmot
Department of Epidemiology and Public Health, University College London, London, UK

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_23,

41.1 Introduction

Social epidemiology has been defined as the branch of epidemiology that studies the social distribution and social determinants of health (Berkman and Kawachi 2000). As all aspects of human life are inextricably bound within the context of social relations, every conceivable epidemiological exposure is related to social factors. In this broad sense, all epidemiology is social epidemiology (Kaufman and Cooper 1999) with perhaps the latter discipline making explicit the analysis of the social determinants of health. However, the term "social" has also been used to contrast with the "individual" and especially individualist theories of society. And so for some, social epidemiology and its social theories of disease distribution stand in contrast to individualistic epidemiology, which relies on individualistic theories of disease causation (Krieger 2001a).

The idea that social conditions influence health is not new. Chadwick (Flinn 1965) wrote about the insanitary conditions of the working classes and how overcrowding, damp, and filth contributed to their lower life expectancy. Durkheim (1897) wrote about how social norms and conditions affect risks of suicide in the population. Social epidemiology builds and expands on this literature by posing new research questions, utilizing new research methods and influencing government policy agenda. The rest of this chapter will discuss each of these three developments in social epidemiology.

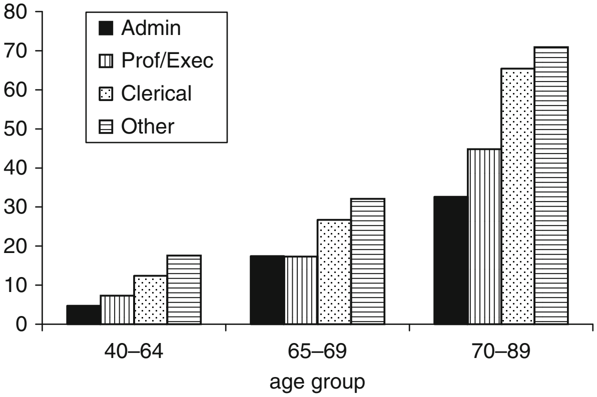
41.2 Research Questions

41.2.1 The Social Determinants of Health

If the social environment is an important cause of health, this is likely to be manifested as social inequalities in health. People from better social environments with greater access to socioeconomic resources are likely to have better health. Supporting this view, social inequalities in health have been documented for most countries, for most causes of deaths and diseases, and in most age-groups. People from lower socioeconomic backgrounds are more likely to be unhealthier and have lower life expectancies, even in the richest countries. In Fig. 41.1, from the first Whitehall study on the health of civil servants in the United Kingdom (Marmot and Shipley 1996), men in the lowest, office support employment grades have mortality rates four times that of men in the highest administrative grade in the youngest age-group. This difference in mortality between hierarchies in the civil service remains even after retirement among men in the oldest age-group. What remains unclear are the pathways leading from the social structure to health or the social determinants of health.

There have been a number of attempts to delineate the pathways underlying the social determinants of health. One such example is illustrated in Fig. 41.2 (from the Commission on Social Determinants of Health (CSDH) 2008). The socioeconomic

Fig. 41.1 All-cause mortality per 1,000 person-years by employment grade: Whitehall men, 25-year follow up (From Marmot and Shipley 1996)



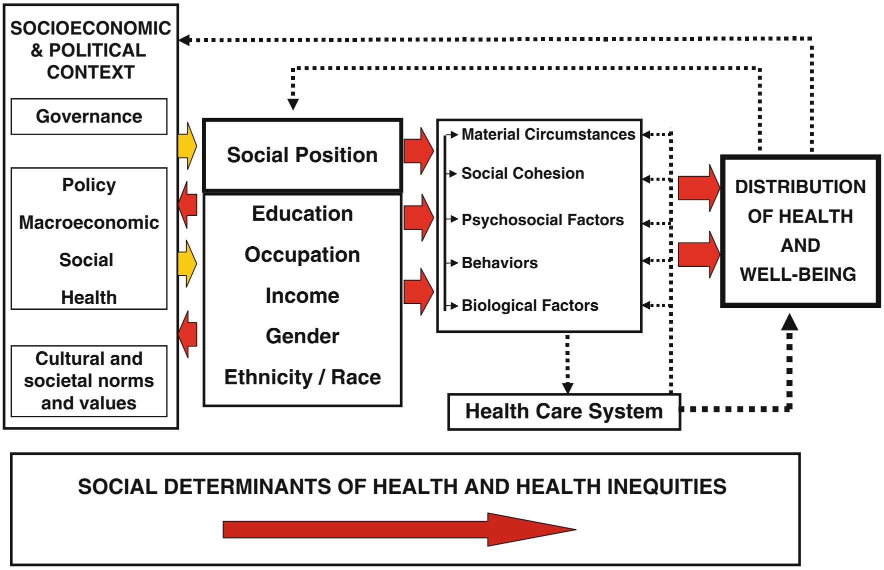


Fig. 41.2 Conceptual framework on social determinants of health. Commission on Social Determinants of Health Final Report (CSDH 2008)

and political context influences the structural determinants of inequalities (social class, gender, race, education) which in turn affect the conditions in which people live, work, and grow, resulting in an unequal distribution of health in the population. In this conceptualization, taking action on the social determinants of health requires action at different levels. On an individual and household level, these include taking action on material and stressful living and working conditions, changing unhealthy behaviors and increasing access to quality health care. On a social level, these also include taking action on the macropolitical and macroeconomic processes that result in unequal employment and educational opportunities and gender and racial discrimination.

There has been relatively little testing of these pathways between social structure and health, primarily because to date, there have been little data available to test all these pathways. However, a few studies have examined some of these pathways and their contribution toward understanding social inequalities in health. The rest of this section will highlight some of the research on the search for the social determinants of health.

41.2.2 Health Behaviors

People from lower socioeconomic groups are more likely to smoke, drink alcohol excessively, and have less physical exercise and unhealthier diets. It is likely that such unhealthy behaviors form part of the pathways underlying social inequalities in health. Poor people in the UK are less likely than those who are well-off to eat a good diet, more likely to have a sedentary lifestyle, more likely to be obese, and more likely to be regularly drunk (Fig. 41.3, from Colhoun and Prescott-Clarke 1996).

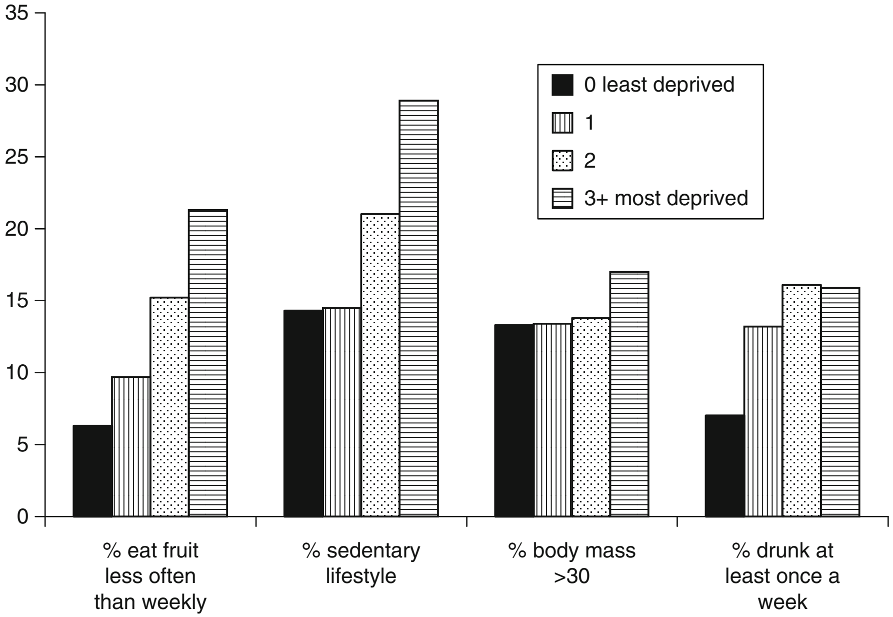


Fig. 41.3 Distribution of some health behaviors among men by level of socioeconomic deprivation (From Colhoun and Prescott-Clarke 1996)

Some studies have analyzed the contribution of such health behaviors to explaining the social gradient in health and have found that a substantial social gradient in health still remains even after adjusting for such (un)healthy lifestyles (Marmot et al. 1978). So there may be other social determinants of health not directly related to health behaviors, as suggested by the pathways through work and material factors in Fig. 41.2. Some evidence of these other pathways is discussed in Sects. 41.2.3–41.2.7.

Lynch et al. (1997) argue that we still need to understand why poor people behave poorly. Without some understanding of how the social environment influences behavior (through, e.g., social norms or environments which may be health damaging or health promoting such as workplace restrictions on smoking or stressful environments for which smoking may be an effective, albeit temporary coping strategy), interventions to modify behavioral risk factors may not be successful.

41.2.3 Material, Economic and Political Determinants of Health

The link between health and material or socioeconomic circumstances has been observed at least since mid-nineteenth-century Britain, if not earlier. Chadwick (Flinn 1965) wrote about how overcrowding and damp and filthy living conditions contributed to the lower life expectancy of working class men. In 1848, partly through fear of cholera and partly through pressure from Chadwick, the British parliament passed the first Public Health Act. This, in addition to the pioneering work of the epidemiologist Snow (1855), set in motion the public health movement in nineteenth-century Britain which saw improvements in housing, sewage and drainage, water supply, and contagious diseases and provided Britain with the most extensive public health system in the world.

It has also been argued that much of the decrease in the mortality rate in the nineteenth and early twentieth century was primarily due to better nutrition in the population which led to increased host resistance to opportunistic infections (McKeown 1979). The driver behind better diets in the general population can be traced to economic growth which made nutritious foods more easily affordable by most of the population. Others, like Szreter (1988), argue that the public health movement of the mid nineteenth century in the UK also played an important role in combating deaths due to infectious disease. It is likely that a combination of macroeconomic factors (economic growth) and public policies (public health measures) led to the overall decreases in mortality rates due to infectious diseases and increases in life expectancy.

In twentieth-century industrialized societies, infectious diseases played an increasingly smaller role in causing deaths, while chronic diseases such as heart disease and cancers caused the majority of deaths. Although people from poorer social classes are more susceptible to such chronic diseases (repeating the patterns of infectious diseases like cholera in nineteenth-century Britain), the mechanisms underlying this social patterning of chronic diseases are not easy to specify. A single infectious agent such as a bacterial agent which thrives in unhygienic circumstances is unlikely to account for why poorer, less advantaged people have more heart attacks.

Some authors argue that, even today, economic and political processes are the fundamental determinants of health and disease (Coburn 2000; Navarro and Shi 2001). Determinants of health can be analyzed in terms of who benefits from specific government policies and practices. Economic and political institutions and decisions that create, enforce, and perpetuate social inequality also create and maintain social inequalities in health. For example, neoliberal (market oriented) policies which favor the dismantling of the welfare state may help to widen existing social inequalities in health. Navarro and Shi (2001) found that countries with more economic and social redistributive policies (Sweden, Finland, Norway, Denmark, and Austria) were more successful in improving the health of their populations (reducing their infant mortality rate). In contrast, neoliberal countries (Canada, United States, United Kingdom, Ireland) where the market reigns supreme and the welfare state is the weakest had the lowest rates of improvements in the infant mortality rates. The substantial decline in life expectancy in Russia in the 1990s has been linked to its transition to a neoliberal economy (Walberg et al. 1998).

41.2.4 Life Course

The idea that a person’s experiences over a lifetime can have a cumulative effect on their health is a central idea within social epidemiology. The study of long-term effects of physical and social exposures during gestation, childhood, adolescence, young adulthood, and later adult life on the risk of chronic disease has been defined as a life course approach to chronic disease epidemiology (Ben-Shlomo and Kuh 2002; see also chapter ►Life Course Epidemiology of this handbook). Such studies include biological, behavioral, and psychosocial pathways that operate over an individual’s life course, as well as across generations, to influence the course of chronic disease. However, it is only in fairly recent years that adequate data and appropriate statistical methods have been made available to test the hypotheses associated with a developmental and life course perspective.

A systematic review on the evidence of different ways in which socioeconomic position over the life course could affect cardiovascular health described four common life course models (Pollitt et al. 2005). The first is a latency model (or “critical period”) of early socioeconomic disadvantage which affects cardiovascular health later on in adulthood independent of intervening socioeconomic circumstances (Power and Hertzman 1997). Barker found evidence that birthweight and other indicators of fetal growth in the newborn are related to fibrinogen and insulin resistance 50 years later (Barker 1991, 1997). Low birthweight is also associated with poorer childhood health which some researchers have linked to lower social position in adulthood (Illsley 1986). Others suggest that there are “sensitive periods” during a child’s development when exposure to socioeconomic advantage or disadvantage is critical for later life health (Hertzman et al. 2001).

The second is the pathway model (Kuh and Ben-Shlomo 1997) in which exposure to socioeconomic disadvantage influences later life socioeconomic position resulting in “social chains of risk.” For example, childhood circumstances may not affect adult risk of ill health and disease directly. It is possible that parental social class and educational qualifications are important because they help to determine the social circumstances in which the offspring lives and works in adult life, and it is these circumstances that give rise to social inequalities in disease. Some studies have found the relationship of education on adult health can be explained in terms of occupational class and income (Davey Smith et al. 1998).

The third is a social mobility model in which upward or downward social mobility affects health, although there is disagreement over whether such social mobility results in wider (West 1991) or narrower (Bartley and Plewis 1997) social inequalities in health. The fourth is a cumulative model in which the accumulation (but not the interaction) of socioeconomic disadvantage throughout the life course affects health (Davey Smith and Lynch 2004), but the timing of exposure to disadvantage (whether in childhood or in adulthood e.g.) is not important.

These four models of life course socioeconomic factors overlap conceptually. For example, the pathway model is a special case of the accumulation model in which early socioeconomic position does not have a direct effect on later life health independent of later life socioeconomic pathways. Similarly, the social mobility model is analogous to a critical period model with later life effect modification, where the exposure to early life disadvantage is either enhanced or diminished by later life socioeconomic position (Mishra et al. 2009). The review by Pollitt et al. (2005) found moderate support for the role of socioeconomic disadvantage in childhood as a critical period for risk of cardiovascular disease, little support for the effect of social mobility, and consistent support for the accumulation model of socioeconomic disadvantage. From a social epidemiological perspective, the lack of evidence for the social mobility model is perhaps a little disappointing, as evidence for these models could potentially provide a way of reducing the negative health consequences of exposure to early life disadvantage. However, the Pollitt review also mentions considerable methodological issues in the analyses of all life course models and advocates the use of multiple socioeconomic measures and multiple life course analytical approaches in understanding chronic disease epidemiology.

41.2.5 Social Biology

Human beings are both social and biological, and understanding the interaction between the two is crucial to understanding the social determinants of health. The biological processes that underlie the social determinants of health make explicit the pathways from psychosocial factors to biological responses. Psychosocial factors may affect health in two distinct ways – they may directly cause biological changes which predispose to disease, or they may, indirectly, influence behaviors such as smoking and diet, which in turn affects health (Brunner 2000).

The direct effect of psychosocial factors on biology may be through the experience of chronic stresses which in turn modify neuroendocrine and physiological functioning (Selye 1956). Humans are adapted to meet the challenge of short-term threats. However, frequent and prolonged activation of the fight-or-flight response or defense reaction appear to be maladapted (Sapolsky 1993). The main axes of the neuroendocrine response appear to be the sympathoadrenal and hypothalamic-pituitary-adrenal (HPA) systems (Brunner 2000). The former, the sympathoadrenal system, is characterized by the rapid release of adrenaline from the adrenal medulla and noradrenaline from the sympathetic nerve endings, which produces, among other things, cognitive arousal, raised blood pressure, and glucose mobilization. There is evidence of wide variations between individuals in the size and duration of these endocrine responses attributed to individual differences in psychological coping resources (Grossman 1991). The HPA system involves cortisol release from the adrenal cortex. Like the sympathoadrenal system, functioning of the HPA axis also appears to be conditioned by psychosocial factors (Hellhammer et al. 1997). Lower social position is associated with prolonged elevations or cortisol release or blunted responses from a raised baseline (Suomi 1997). These patterns of cortisol secretion differ from the normal sharp response and rapid return to a low baseline. A comparison of Swedish and Lithuanian men given a stress test revealed higher morning cortisols and blunted reactivity among the low-income group drawn from the higher coronary risk Lithuanian population (Kristenson et al. 1998). Low socioeconomic position among British civil servants (the Whitehall II cohort) is associated with blunted cortisol reactivity in terms of a flatter slope in the diurnal cortisol profile (Kumari et al. 2010).

There is some evidence for the hypothesis that psychosocial factors directly affect neuroendocrine mechanisms which result in social inequalities in coronary heart disease. Hostility and anxiety have been linked with reduced heart rate variability (HRV) which refers to the beat-to-beat alterations in heart rate (Hemingway et al. 1998). HRV appears to be sensitive and responsive to acute stress as well as a marker of cumulative wear and tear. HRV has been shown to decline with the aging process which has been attributed to a decrease in efferent vagal tone and reduced beta-adrenergic responsiveness. By contrast, regular physical activity (which slows down the aging process) has been shown to raise HRV, presumably by increasing vagal tone.

The metabolic syndrome is a well-known precursor state to coronary heart disease (CHD) and is linked with increased risk of type 2 diabetes. The main components of the metabolic syndrome are impaired glucose tolerance, insulin resistance, and disturbances of lipoprotein metabolism characterized by raised serum triglycerides and low HDL cholesterol (Seidell et al. 1990). Although the link between the metabolic syndrome and CHD is well established, the association between psychosocial factors and the metabolic syndrome is less certain. Central obesity and other components of the metabolic syndrome are consistently related to low socioeconomic position in industrialized countries (Brunner et al. 1993; Kaplan and Keil 1993). It is possible that chronic psychosocial stresses result (directly) in the metabolic syndrome pattern of abnormalities through the activation of the HPA axis. Increased HPA activity results in redistribution of body fat, leading to central obesity, hypertension, and type 2 diabetes, as found in Cushing's syndrome (Howlet et al. 1985). The alternative explanation is that psychosocial stresses lead to unhealthy behaviors (smoking, inappropriate diets).

However, in the Whitehall II cohort, adjusting for health behaviors did not change the social gradient in the metabolic syndrome, suggesting a direct neuroendocrine effect (Brunner et al. 1997).

Infectious disease may also contribute to social differences in morbidity. *Helicobacter pylori* infection, acquired in childhood, is linked with deprivation and overcrowded housing and may produce long-term low-level systemic inflammatory responses which enhance atherogenesis. In Whitehall II, lower employment grade in the civil service is linked to raised fibrinogen (Brunner et al. 1995) and higher concentrations of CRP and IL-6 (Gimeno et al. 2007).

41.2.5.1 Epigenetic Changes in Relation to Social Conditions

While it is unlikely that there are social position differences in genetic structure (Holtzman 2002), how genes are expressed may be related to social conditions. Whether a gene makes its product (gene expression) can be influenced by "epigenetic" processes, so called because the change in gene expression occurs without a change in DNA structure. In many cases, the epigenetic process is the methylation of DNA – the DNA molecule is chemically modified by the addition of methyl residues which confer a change in gene expression. Pertinent to the latency life course model, one of the best-known examples of this phenomenon in humans is the demonstration of persistent epigenetic differences associated with famine in utero. In this example, individuals aged 58 on average who were exposed to famine in utero in early pregnancy had different rates of methylation of a gene involved in growth (IGF-2) compared to non-exposed siblings (Heijmans et al. 2008).

After birth, studies focus on the maternal/neonatal interactions and their association with long-term changes in gene expression. In animal studies, experiments have focussed on long-term changes in neuroendocrine function, specifically in the hypothalamic-pituitary-adrenal axis and a receptor within this system, the glucocorticoid receptor. For example, a number of studies have demonstrated that the epigenetic status of the glucocorticoid receptor gene promoter is regulated by parental care during early postnatal development in rats (Meaney and Szyf 2005).

Data in humans are currently lacking, but early evidence may suggest that the social environment early in life has a long-lasting biological and health impact via epigenetic marking of specific genes. For example, methylation profile of the glucocorticoid receptor is different in those that have committed suicide associated with childhood maltreatment relative to controls (McGowan et al. 2009). DNA methylation is potentially reversible (Weaver et al. 2004), making this process subject to a number of life course processes; however, these remain to be determined in human studies.

41.2.6 Ecological Perspectives

In the UK and elsewhere, there are marked differences in health between areas. People living in areas with higher levels of poverty have poorer health on average and lower average life expectancy. However, explanations for these area differences in health remain debatable. Some argue that excess mortality in deprived areas can be wholly explained by the concentration of poorer people in those areas (Slogett and Joshi 1994; Duncan et al. 1993). In other words, the compositional or aggregate effect of poor individuals (each of whom has lower than average life expectancy) in an area explains the lower average life expectancy for the area. Others argue that such compositional effects cannot entirely explain area differences in health (Diez-Roux et al. 1997). They point out that even after adjusting for the composition of individuals living in an area (such as their income and wealth levels), significant area differences in health remain. They argue that there may be contextual or ecological reasons for area differences in health, explanations which may not be reduced to an aggregation of the individuals living in the area. For example, particular characteristics of an area such as its pollution levels or its lack of medical services may have an impact on the health of everyone living in that area (Macintyre et al. 1993). Research from the United States has found that states with lower levels of trust have higher rates of violent crime, including homicide (Kawachi et al. 1999a). Such contextual effects may also interact with an individual's characteristics, and this combined interaction may alter their risk of disease. For example, the lack of medical services may have a greater impact on the health of poor people living in an area compared to richer people who may have the resources to travel or access medical services outside their local area. Such ecological or contextual characteristics clearly form part of the social determinants of health and may play some role in explaining social inequalities in (individual) health.

Ecological approaches were disfavored for many years in social epidemiology (Macintyre and Ellaway 2000). Although public health practitioners in the nineteenth and early twentieth centuries focused on dealing with health damaging and promoting environments such as sewage, clean water, housing, and physical working conditions, the decline in infectious diseases led to less emphasis being placed on such ecological factors. The rediscovery of social inequalities in health toward the end of the twentieth century focussed primarily on the role of individual health-risk factors such as behaviors, low income, lack of employment and education, and a relative neglect of contextual or environmental determinants of health. This neglect has been explicitly addressed in the most recent literature with multilevel analyses that explicitly take into account compositional and contextual social factors that affect health (Macintyre and Ellaway 2000).

41.2.7 General Susceptibility to Disease

According to the general susceptibility hypothesis (Syme and Berkman 1976), social factors influence disease by creating a vulnerability or susceptibility to disease in general rather than to any specific disorder. This idea was built on the observation that many social conditions are linked to a broad range of diseases. While behavioral, environmental, biological and genetic factors influence specific diseases, these factors may interact with socially stressful conditions in the development of these diseases, resulting in illness and early mortality.

As discussed above, research from social biology shows that some stressful experiences activate multiple hormones, affecting multiple systems and potentially producing wide-ranging organ damage. The cumulative experience of stress may affect a variety of chronic and infectious diseases through neuroendocrine-mediated biological pathways. There are a number of different sources of stressful experiences, some of which are discussed below. The linking of such stressful experiences (often measured using psychological concepts) to wider social circumstances has been called a psychosocial approach to understanding the social determinants of health.

41.2.7.1 Social Support

The effect of social support and social networks on health has been researched (at least) since the late nineteenth century when Durkheim investigated the links between social integration and suicide. He explained suicide in terms of social dynamics, arguing that suicide is not an isolated individual tragedy but a reflection of social conditions such as the lack of attachment and regulation in society. Attachment is also a core concept for Bowlby (1969) who argued that marriage is the adult equivalent of childhood attachment between mother and child. Secure attachment, whether in terms of parent-child or marital relationships, provides for successful and healthy development. Men who have never married or have recently divorced have a significantly greater risk of dying from both cardiovascular and non-cardiovascular diseases than married men (Ebrahim et al. 1995). Married women are generally healthier than unmarried women as well, although the health benefits of marriage may not be particularly strong for employed women (Waldron et al. 1996).

Throughout the 1970s and 1980s, a series of studies appeared which consistently showed that the lack of social ties or social networks predicted mortality from almost every cause of death (Berkman 1995). Social ties and networks were measured in terms of numbers of close friends and relatives, marital status, and membership in religious or voluntary associations. Since then, studies have gone on to focus on the provision of social support rather than on the elaboration of the structural aspects of social networks. Not all social ties or networks are supportive, and there is variation in the type, frequency, and extent of support provided. Social support, in theory, can be divided into emotional support (usually provided by a confidant or intimate other), instrumental support (or help in kind, money, or labor), appraisal support (help in decision making), and informational support (provision of advice or information). Lack of emotional support has been linked to early cardiovascular disease mortality among both men and women and younger and older people (Berkman et al. 1992). Other studies have found that social integration, particularly operating though emotional support, influence recovery from strokes (Berkman and Glass 2000). However, this is contradicted by evidence from a randomized control trial aimed at improving social support and reducing depression in postmyocardial infarction patients. The ENRICHD study found no difference in survival between the control and intervention (which had increased social support and lower depression) groups (Berkman et al. 2003).

41.2.7.2 Social Disorganization

Social scientists have puzzled over why some societies seem to prosper, possess effective political institutions, and have better health outcomes compared to other societies. One of the hypotheses that has been proposed to explain this difference between societies is the amount of social capital or cohesion (and its converse – social disorganization) in a particular society (Coleman 1988; Putnam 2000). Social cohesion refers to the extent of connectedness and solidarity among groups in society. A cohesive society has greater amounts of social capital (higher levels of interpersonal trust, reciprocity, and mutual aid) than a disorganized society. There is emerging evidence that greater social capital is linked to lower mortality rates as well as better self-rated health (Kawachi and Berkman 2000). As mentioned in Sect. 41.2.6, states in the United States with lower levels of trust have higher homicide rates (Kawachi et al. 1999a). Even after adjusting for individual risk factors for poor self-rated health (e.g., low income, low education, smoking, obesity, lack of access to health care), individuals living in US states with low social capital were at increased risk of poor self-rated health (Kawachi et al. 1999b). Such results suggest that there are contextual explanations for area differences in health, as discussed in the section on ecological perspectives.

Social capital may be linked to health through a number of different mechanisms. We have already discussed two types of explanations for understanding area differences in health – compositional and contextual explanations. Socially isolated individuals (not having contacts with friends or relatives, not belonging to any groups) are more likely to be living in communities with lower social capital, so the association between social capital and health may be the compositional effect of the aggregation of socially isolated individuals. However, there may be other pathways by which social capital affects health (Kawachi and Berkman 2000):

1. Through health-related behaviors. Social capital may influence the health behaviors of neighborhood residents by exerting social control over deviant behaviors such as adolescent smoking, drinking, and drug abuse.
2. Through access to services. Socially cohesive neighborhoods are more successful at organization access to services such as transport, health services, and recreational facilities
3. Through psychosocial processes. Socially disorganized neighborhoods with low social capital could have higher levels of fear of crime and other stressors which could negatively impact on the residents' health.

41.2.7.3 Work Stress

One of the more established results in epidemiology has been the link between physical working conditions and health. Reports on occupational health have highlighted the link between emphysema and other lung disease with coal mining, musculoskeletal disorders, and accidents with certain types of manual work. In recent years, there has been increased research on work-related stress and how that affects both physical and mental health.

There are two dominant models of work stress in the literature. The first, the job strain model is based on the concepts of job control and demands (Karasek et al. 1981). Workers with low levels of job control and high levels of demand are said to have high levels of job strain (or work stress). In addition, workers with job strain who have unsupportive managers or colleagues (isolated job strain or iso-strain) have particularly high levels of work stress. Job control (or decision latitude) consists of whether or not workers are able to utilize and develop skills (skill discretion) and their authority over decisions. Job demands consist of qualitative emotional demands as well as quantitative demands specifying output per unit of time. Prolonged and repeated exposure to job strain is hypothesized to increase sympathoadrenal arousal and decrease the body's ability to restore and repair tissues which in turn affects health (Chandola et al. 2008). Civil servants in the UK with greater duration of exposure to job strain and iso-strain have higher risks of cardiovascular disease (Chandola et al. 2006, 2008).

The other model of work stress, the effort-reward imbalance model (Siegrist 1996), hypothesizes that the degree to which workers are rewarded for their efforts is crucial for their health. When a high degree of effort does not meet a high degree of reward, emotional tensions arise and the risk of illness increases. Effort is the individual's response to their job demands, and this response may be extrinsic effort (referring to the individual's effort to cope with external job demands) and intrinsic effort (referring to the individual's drive to fulfill their goals). Reward can be measured through financial rewards, self-esteem, and social control. While there is some overlap between the job strain and effort reward imbalance models, the former is entirely focussed on the organization of the structure of work, while the latter includes the individual's way or coping methods of handling difficulties (through the concept of intrinsic effort).

There is some evidence that both models of work stress contribute independently of one another to predicting coronary heart disease events (Bosma et al. 1998). The cumulative adverse health impact of low job control and effort-reward imbalance indicates that both job stress factors provide supplementary information on the relevant stressors in the psychosocial work environment.

41.2.7.4 Unemployment and Job Loss

There has been considerable research into the effects of unemployment and job loss on health. However, this is an area of research that is particularly sensitive to the claims of "health selection," that the reason why unemployment is associated with ill health is because ill health selects people out of employment. The reverse argument is that a disadvantaged socioeconomic position has an effect on a stable job career (and the risk of unemployment) as well as health. It is therefore important to disentangle the causal narrative in studies about unemployment and health and find out which comes first.

The evidence on unemployment and health supports both the social causation and health selection interpretations. In a review of the effect of unemployment on health, Kasl and Jones (2000) summarized the evidence as follows:

1. Unemployment is associated with a 20–30% excess in all-cause mortality in most studies.
2. There is some evidence of the impact of unemployment on physical morbidity but with results that are more difficult to interpret.
3. Unemployment is linked to biological indicators of stress reactivity.
4. Unemployment is associated with behavioral and lifestyle risk factors, although the direction of causality is hard to disentangle.
5. Unemployment clearly increases psychological distress.
6. Threatened job loss (job insecurity) is associated with physical and psychological morbidity and cardiovascular risk. The anticipation of job loss affects health even before changes in employment status (Ferrie et al. 1995).

41.2.7.5 Depression and Affective States

Depression is one of the most common psychiatric problems and is also common in patients with chronic medical conditions. Some depressive episodes are brought upon by physical illness, but many depressive patients have depressive episodes long before they develop any physical symptoms of illness. Furthermore, depression may alter the course and outcome of physical illness (Carney and Feedland 2000).

Depression has been associated with immunological dysfunction. Patients with major depression have been found to have blunted natural killer cell activity (Maes et al. 1994), increasing their risk for many acute and chronic illnesses. There is also some evidence that depression may play a causal role in the development of heart disease. There is some evidence of a social gradient in depression in a healthy, working population – it appears to be more common among those from poorer, more disadvantaged social positions (Stansfeld et al. 1998) and may originate from their lower control over aspects of their work and home environment (Griffin et al. 2002).

Another set of psychological pathways by which social conditions may affect health is through emotions and the physiological, cognitive, and behavioral responses they evoke. Emotions may be transitory states brought on by specific situations or traits, i.e., stable and general dispositions to experience particular emotions (Spielberger and Krasner 1988). Much of the research on emotion and health has been carried out in relation to coronary heart disease. Much of this literature has focussed on type A behavior (which includes a free-floating but well-rationalized hostility, hyperaggressiveness, and a sense of time urgency), chronic anger and hostility, anxiety, and a mixture of emotions associated with depression including hopelessness, loneliness, guilt, and shame (Kubzansky and Kawachi 2000).

There is some evidence of a social patterning of emotions (Bradburn 1969; Mroczek and Kolarz 1998). Kemper (1993) suggests that many emotions are responses to power and status differentials embedded within social situations. Potentially stressful events can be associated with a variety of different emotions. Emotions can be considered as products of stress as well as mediators of its effects, thus representing a crucial link in the chain of causation from social stressors to individual biological responses (Spielberger and Krasner 1988). Evidence from animal studies suggests that additional to hypothalamic control of the stress response, areas of the brain involved with emotional or affective responses such as the limbic system also play a major role in stress responses (Menzhaghi et al. 1993) and adaptation to the stress response (Sapolsky et al. 1986).

41.3 Research Methods

41.3.1 Applying a Population Perspective

Rose (1992) proposed that an individual's risk of illness cannot be considered in isolation from the risk of disease of the population to which they belong. For example, the distribution of cholesterol levels in the Finnish population is shifted to the right of the Japanese distribution – on average, Finnish people have higher cholesterol levels than Japanese people. The level of "normal" cholesterol for the Finnish population would be "abnormal" for the Japanese population and would be a risk factor for CHD in the latter population. Applying the population perspective into epidemiological research means asking "why does this population have this particular distribution of risk factors," in addition to asking "why did a particular individual get sick?" (Berkman and Kawachi 2000). Answering the second question has been the focus of clinical medicine, while answering the first question is the key to the largest improvements in the health of the population as it focuses attention on the majority of cases of illness within the bulk of the population. Medical care can prolong survival after some serious diseases, but the social and economic conditions that affect whether people become ill are more important for health gains in the population as a whole.

41.3.2 Better Measures of Social Exposures

There is no simple relationship between social-structural conditions such as income distribution and welfare state regimes on the one hand and health inequalities in the population on the other. The different pathways by which social factors can have an effect on different health outcomes implies that there is no single measure of socioeconomic position that can represent all of these social dimensions. However, it is common for epidemiological studies to state they have taken account of the confounding effects of socioeconomic position by adjusting for a single crude measure of socioeconomic position (SEP). The possibility that the measure of SEP does not adequately measure the different dimensions of SEP in the population, at different stages of the life course, is seldom considered in standard epidemiological analyses.

Measures of socioeconomic position are not interchangeable. The Commission on Social Determinants of Health (Kelly et al. 2007) identified the structural determinants of health inequalities as the social processes by which inequalities are generated and reinforced; these include social (occupational) class, power, status, and education.

Many epidemiological studies only measure a single measure of socioeconomic position, most commonly education. Although education is relatively easy to measure (either in terms of years spent in education or in terms of highest qualifications obtained), it has the major limitations. Education is largely invariant over the adult life course, and it is hard to conceptualize of downward social mobility in terms of education (hence, it is not really a measure of "position"). Education predicts but does not directly measure adult socioeconomic position, thus resulting in biased estimates of adult socioeconomic position in many epidemiological studies.

Income, wealth, and assets are good indicators of a person's position in the labor market as well as their material standards of living. Higher levels of income and wealth enable greater access to and greater consumption of health promoting resources and services. Accurate measurements of income require detailed information from multiple income streams from all individuals in the same household which is not always possible in surveys. Data on income are often sensitive. People are not willing to disclose how much they earn, resulting in a large proportion of non-response or missing data from surveys on income. Many studies use proxy measures of material living standards such as entitlements to social security benefits, possession of consumer goods and assets, wealth, car ownership and house ownership, (or "consumption" measures) which are less prone to non-response bias. Wealth represents the total value of a person's or household's assets. These can be a wide variety of assets and require valuation of non-monetary assets such as land and housing, which are difficult to estimate and measure. Some argue that consumption expenditure is a more accurate representative of long-term economic status than income (Friedman 1957). This is because income is subjective to greater fluctuations, whereas individuals and households may base their consumption decisions on their planned and anticipated ("permanent") income rather than their current income levels.

In the developed world, some authors argue that occupational class is the key measure of SEP (Rose and Pevalin 2002). A review of measurement of social circumstances concluded that a classification based on occupation was essential to understanding social inequality (Rose and Pevalin 2002). Some argue that occupation is the key variable in the accumulation of advantages and disadvantages over a person's life course. Occupation can be regarded as the means by which a person's principal resource (education) is converted into an important reward (income). As occupation links these two sets of advantages, it is a greater measure of accumulated advantages than either one alone.

Different countries often have national occupational class schemas which makes cross-country comparisons difficult. However, the International Labor Organization has developed the International Standard Classification of Occupations based on two main concepts: the kind of work performed and the skill involved. Another limitation of occupational class is the lack of comprehensive population coverage – vulnerable groups such as children, women, and the elderly often have to be classified by proxy measures of social class in which case, these social class measures no longer reflect differences in employment conditions, but rather they reflect differences between households in membership of particular occupations.

41.3.3 Better Measures of Health

The concept of health is multidimensional (the WHO definition states that health is a state of complete physical, mental, and social well-being), including hard to measure concepts like quality of life. Research in social epidemiology does not just focus on clinically measured disease outcomes because the absence of disease is not sufficient for health. Rather, one of the main focuses of social epidemiological research is the use of health-related quality of life measures as valid measures of health outcomes (Fitzpatrick et al. 1992). Population mortality statistics tell us little about the health of general populations in developed countries. The use of standardized health-related quality of life measures in different countries (Ware and Gandek 1998) enables international comparisons of physical, mental, and social well-being.

Subjective health status covers a wide variety of areas, including role functioning (e.g., the ability to perform domestic and work tasks), the degree of social and community interaction, psychological well-being, pain, tiredness, and satisfaction with life (Bowling 2003). Health-related quality of life has come to mean a combination of subjectively assessed measures of health, including physical function, social function, emotional or mental state, burden of symptoms, and sense of well-being (Coulter 1993). Some of these patient-based measures of health status are reviewed in chapter ▸Health Services Research of this handbook. Concepts of health-related quality of life adjusted measures of life expectancy such as disability-adjusted life years (DALYs) and quality-adjusted life years (QALYs) are also reviewed in this chapter.

41.3.4 Better Measures of the Association Between the Social Structure and Health

Social epidemiology has typically relied on observational studies to describe the effect of social factors on health such as social inequalities in health. As associations described in observational studies may be subject to confounding bias, especially from unmeasured confounders, it is hard to draw causal inference from most social epidemiological studies. There are many examples of how results from observational studies in social epidemiology have not been replicated in randomized control trials (Berkman 2008), although there have also been examples where randomized control trials have replicated results from observational social epidemiology studies (Howden-Chapman et al. 2007).

Methodological advances in social epidemiology in using propensity score matching (Oakes and Johnson 2006), natural experiments and instrumental variables (Glymour 2006), and controlled community trials (Hannan 2006) offer new ways of building causal and explanatory models. Standard multiple regression procedures do not necessarily advance a causal understanding of the social determinants of health. For example, as different measures of the social structure may have different pathways to different health outcomes, the reduction of such differences into a single regression model may obscure rather than elucidate the pathways underlying the social determinants of health. Furthermore, different dimensions of the social structure may influence people's health at different time points of the life course. For example, in industrialized societies, the period of the life course when compulsory education is completed may be a crucial time for the health of the population, not because young adults are at a particular high risk of disease or illness at that stage in life but because educational qualifications are a strong determinant of social position in later adult life which in turn appear to be strongly linked to health outcomes later on in life. It is important to take account of the life course temporal and causal ordering of the various measures of social position and use methods that make explicit the various underlying causal pathways between different measures of social position and health.

41.3.5 Analyzing Population Surveys, Birth Cohorts

One of the defining characteristics of research methods in social epidemiology is the use of population representative sample surveys in analyzing the social determinants of health. Research in social epidemiology tends to use non-experimental observational studies, both cross-sectional and longitudinal. All observational studies suffer from problems of causality – it is hard to determine and separate out cause from effect (cf. chapters ▶Basic Concepts and ▶Confounding and Interaction of this handbook). This drawback has necessitated the use and development of complex study designs and analytical methods to disentangle the causal pathways underlying the social determinants of health.

Studies with good methodological designs (e.g., Ferri et al. 2003) in social epidemiology tend to rely on data from large-scale population representative sample surveys because of the complexity of the social structure and the different pathways to health. The representativeness of data is crucial in order to apply a population perspective in social epidemiological research. Smaller-scale samples may not be representative of the broader population.

Birth cohort studies are a special type of such large-scale population representative samples which incorporates a life course approach to epidemiology. The UK has taken a prominent role in the development of such longitudinal studies. The British birth cohort studies of those born in one week of 1946, 1958, and 1970 link data from one part of the life course (from birth onward) to another (childhood, adolescence, adulthood) for a large number of individuals. Comparisons between different birth cohort studies enable the disentangling of age, period, and cohort effects, which could be problematic when analyzing most cross-sectional and even longitudinal sample surveys. For example, in the book, “Changing Lives, Changing Britain” (Ferri et al. 2003), *cohort* effects that might be attributed to socioeconomic change impacting differentially on people born at different times can be differentiated from *age* differences reflecting the different changes between the stages of life, which in turn can be differentiated from the prevailing socioeconomic context at the time of data collection – the *period* effect. Such analyses of this unique set of longitudinal data, incorporating a life course perspective, are very promising for future research into social epidemiology.

41.4 Setting Government Policy Agenda

One of the goals of epidemiology has always been to use what we learn to improve public health. The science of social epidemiology has repeatedly shown evidence that social conditions are a major determinant of health. However, the translation of the research findings of social epidemiology into public policy has not been straightforward. Unlike results from some branches of epidemiology which can be more easily implemented into government guidelines (such as recommended alcohol intake) or public policy (reduction in smoking prevalence), programs to implement findings from social epidemiology need to take into the account the complexity of the pathways from the social structure to health.

Some authors argue that policy interventions are most effective when they are closest to the root causes of disease (Rothman et al. 1998). Interventions at the upstream, social level may not be as efficient as interventions closer to disease occurrence. So, for example, policy interventions on reducing the social gap in smoking-related diseases should focus on interventions on smoking cessation rather than interventions on the social causes of smoking. Others argue (such as Coburn (2000), mentioned in Sect. 41.2.3) that interventions need to be upstream, at the societal and macroeconomic level, in order to successfully reduce health inequalities.

41.4.1 Reports on Inequalities in Health

The Black Report (DHSS 1980) into inequalities in health in the UK had a number of wide-ranging policy recommendations for reducing such inequalities. However, the lack of implementation of these policies by the British government in the 1980s and early 1990s was due, in part, to a lack of political will and the high cost of these policy recommendations.

The change in government in Britain in 1996 (from Conservative to Labor) paved the way for the publication of the Acheson Report (Acheson 1998) on inequalities in health in 1998 with another list of recommendations for reducing health inequalities. The recognition that inequalities in health are evident in every country led the WHO to commission a report on the social determinants of health (CSDH 2008). The commission was established to support countries and global health partners to address the social factors leading to ill health and health inequalities. This has led to many countries commissioning their own reviews of health inequalities and the social determinants of health (http://www.who.int/social_determinants/thecommission/countrywork).

The very fact that social epidemiology deals with the social structure necessitates policies aimed at changing the social structure as well as the intermediate pathways to health. Such policies are not always easy to specify, detail, and enact. Furthermore, the diffuse ownership of such policies between government departments (such as education, health, and treasury departments) makes their implementation harder. In recognition of the complexities of policies aimed at reducing health inequalities, the UK government set up a crosscutting spending review (across various government departments) on tackling health inequalities. This report (Department of Health 2002) explicitly acknowledges that policies on reducing health inequalities need to be coordinated across a wide range of government departments and bodies (not just the national health service), including local government and health organizations and sets in process the institutional framework for such coordination. An evidence-based strategy for reducing health inequalities that includes policies and interventions that address the social determinants of health inequalities has been commissioned in England: the Strategic Review of Health Inequalities in England Post 2010 (The Marmot Review 2010). However, even with evidence-based policy recommendations, political will is also needed for action on health inequalities.

41.4.2 Collating Evidence for Policies Through Intervention Studies and Cross National Comparative Studies

One of the ways of ensuring appropriate policies for reducing inequalities in health are implemented is by studying the results from intervention studies. However, social epidemiological research does not easily lend itself to intervention studies, mainly because the complexity of the social structure makes it hard to disentangle the pathways to reductions in health inequalities. For example, it is hard to disentangle changes from behavioral change interventions from secular trends in society (Susser 1995). It is also difficult to separate out the influence of secondary support (from support groups organized around behavioral interventions) from the intended influence of the behavioral intervention (Spiegel et al. 1989). Social support interventions have had mixed results partly because as relationships develop and change slowly, the benefits of support interventions may be missed in the short term (Glass 2000).

Another method of analyzing policy recommendations for health inequalities is through international and longitudinal comparisons of health inequalities. Changes in taxation and income redistribution policies within a country may be hypothesized to have an effect on health inequalities. Furthermore, cross national longitudinal comparisons of different tax policies and their effect on health inequalities may be another way of analyzing the effect of policies on health inequalities (Navarro and Shi 2001). However, to date, there has been little research in this area which means that current policies on reducing inequalities in health may not be entirely appropriate or well targeted.

41.5 Conclusions

Perhaps, the major contribution of social epidemiology to epidemiology in general has been in rediscovering and analyzing the role of social factors in producing health and illness. This has primarily come about by the literature on social inequalities in health and, consequently, research into the social determinants of health. The search for the pathways between the social structure and health has led to innovations in longitudinal research methodology. While social epidemiology shares common epidemiological problems of reliance on observational studies and problems in interpreting causality, the incorporation of a life course perspective by analyzing and comparing birth cohort studies holds great promise for future studies. Research into social epidemiology has influenced wide ranging government social policies because of the macrosocietal-level interventions that are needed to reduce inequalities in health.

Although there is some debate over the usefulness of the specialization of social epidemiology within the medical sciences (Zielhus and Kiemeney 2001), others have argued that the overall contribution of social epidemiology toward understanding current and changing distributions of population health have been striking (Krieger 2001b; Muntaner 2001). The interdisciplinary nature of social epidemiology has led to the incorporation of research questions, methods, and policy agendas that have enriched our understanding of the social determinants of health.

Acknowledgements Tarani Chandola is supported by a grant to the Whitehall II study by the Medical Research Council. Michael Marmot is supported by a Medical Research Council Research Professorship.

References

Acheson D (1998) Inequalities in health: report of an independent inquiry. HMSO, London

Barker DJP (1991) The foetal and infant origins of inequalities in health in Britain. J Public Health Med 12(2):64–68

Barker DJP (1997) Fetal nutrition and cardiovascular disease in later life. Br Med Bull 53:96–108

Bartley M, Plewis I (1997) Does health-selective mobility account for socioeconomic differences in health? Evidence from England and Wales, 1971–1991. J Health Soc Behav 38:376–386

Ben-Shlomo Y, Kuh D (2002) A life course approach to chronic disease epidemiology: conceptual models, empirical challenges, and interdisciplinary perspectives. Int J Epidemiol 31:293

Berkman LF (1995) The role of social relations in health promotion. Psychosom Med 57:245–254

Berkman LF (2008) Social epidemiology: social determinants of health in the United States: are we losing ground? Annu Rev Public Health 30:19.1–19.15

Berkman LF, Blumenthal J, Burg M, Carney RM, Catellier D, Cowan MJ, Czajkowski SM, DeBusk R, Hosking J, Jaffe A, Kaufmann PG, Mitchell P, Norman J, Powell LH, Raczynski JM, Schneiderman N (2003) Effects of treating depression and low perceived social support on clinical events after myocardial infarction: the Enhancing Recovery in Coronary Heart Disease Patients (ENRICHD) Randomized Trial. JAMA 289:3106–3116

Berkman LF, Glass T (2000) Social integration, networks and health. In: Berkman LF, Kawachi I (eds) Social epidemiology. Oxford University Press, New York, pp 137–173

Berkman LF, Kawachi I (2000) A historical framework for social epidemiology. In: Berkman LF, Kawachi I (eds) Social epidemiology. Oxford University Press, New York, pp 3–12

Berkman LF, Leo-Summers L, Horwitz RI (1992) Emotional support and survival after myocardial infarction. A prospective, population-based study of the elderly. Ann Intern Med 117: 1003–1009

Bosma H, Peter R, Siegrist J, Marmot MG (1998) Two alternative job stress models and the risk of coronary heart disease. Am J Public Health 88:68–74

Bowlby J (1969) Attachment and loss, Vol. 1: Attachment. The Hogarth Press, London

Bowling A (2003) Measuring health: a review of quality of life measurement scales. Open University Press, Milton Keynes

Bradburn NM (1969) The structure of psychological well-being. ALDINE, Chicago

Brunner EJ (2000) Toward a new social biology. In: Berkman LF, Kawachi I (eds) Social epidemiology. Oxford University Press, New York,

Brunner EJ, Marmot MG, Nanchahal K, Shipley MJ, Stansfeld SA, Juneja M, Alberti KGMM (1997) Social inequality in coronary risk: central obesity and the metabolic syndrome. Evidence from the WII study. Diabetologia 40:1341–1349

Brunner EJ, Mendall MA, Marmot MG (1995) Past or present Helicobacter pylori infection and fibrinogen – a possible link between social class and coronary risk? J Epidemiol Community Health 49:545

Brunner EJ, Nicholson A, Marmot MG (1993) Trends in central obesity and insulin resistance across employment grades: the WII Study. J Epidemiol Community Health 47:404–405

Carney RM, Feedland KE (2000) Depression and mental illness. In: Berkman LF, Kawachi I (eds) Social epidemiology. Oxford University Press, New York, pp 191–212

Chandola T, Britton A, Brunner E, Hemingway H, Malik M, Kumari M, Badrick E, Kivimaki M, Marmot MG (2008) Work stress and coronary heart disease: what are the mechanisms? Eur Heart J 29:640–648

Chandola T, Brunner E, Marmot M (2006) Chronic stress at work and the metabolic syndrome: prospective study. Br Med J 332:521–525

Coburn D (2000) Income inequality, social cohesion and the health status of populations: the role of neo-liberalism. Soc Sci Med 51:135–146

Coleman JS (1988) Social capital in the creation of human capital. Am J Sociol 94:S95–S120

Colhoun H, Prescott-Clarke P (1996) Health survey for England 1994. HMSO, London

Coulter A (1993) Measuring quality of life. In: Kinmouth A, Jones R (eds) Critical reading in general practice. Oxford University Press, Oxford

CSDH 2008 (2008) Closing the gap in a generation. Health equity through action on the social determinants of health. Final report of the Commission on Social Determinants of Health. WHO, Geneva

Davey Smith G, Hart CL, Hole DJ, MacKinnon P, Gillis C, Watt G, Blane D, Hawthorne VM (1998) Education and occupational social class: which is the more important indicator of mortality risk? J Epidemiol Community Health 52:153–160

Davey Smith G, Lynch J (2004) Life course approaches to socioeconomic differentials in health. In: Kuh D, Ben Shlomo Y (eds) A life course approach to chronic disease epidemiology, 2 edn. Oxford University Press, Oxford, pp 77–115

Department of Health (2002) Tackling health inequalities: summary of the 2002 cross-cutting review. Department of Health, London

DHSS (1980) Inequalities in health: Report of a research working group chaired by Sir Douglas Black. DHSS, London

Diez-Roux AV, Nieto FJ, Muntaner C, Tyroler HA, Comstock GW, Shahar E, Cooper LS, Watson RL, Szklo M (1997) Neighborhood environments and coronary heart disease: a multilevel analysis. Am J Epidemiol 146:48–63

Duncan C, Jones K, Moon G (1993) Do places matter: a multilevel analysis of regional variations in health related behaviour in Britain. Soc Sci Med 42:817–830

Durkheim E (1897) Suicide. Free Press, New York

Ebrahim S, Wannamethee G, McCallum A, Walker M, Shaper A (1995) Marital status, change in marital status, and mortality in middle-aged British men. Am J Epidemiol 142:834–842

Ferri E, Brynner J, Wadsworth M (2003) Changing Britain, changing lives. Education Press, London

Ferrie JE, Shipley MJ, Marmot MG, Stansfeld S, Davey Smith G (1995) Health effects of anticipation of job change and non-employment: longitudinal data from the Whitehall II study. Br Med J 311:1264–1269

Fitzpatrick R, Fletcher A, Gore S, Jones D, Speigelhalter D, Cox D (1992) Quality of life measures in health care. I: applications and issues in assessment. Br Med J 305:1074–1077

Flinn, M.W. (1965) Introduction, in Report on the Sanitary Condition of the Labouring Population of Great Britain by Edwin Chadwick 1842 Edited with an introduction by M.W. Flinn. University Press, Edinburgh

Friedman M (1957) A theory of the consumption function. Princeton University Press, Princeton/New Jersey

Gimeno D, Brunner EJ, Lowe GD, Rumley A, Marmot MG, Ferrie JE (2007) Adult socioeconomic position, C-reactive protein and interleukin-6 in the Whitehall II prospective study. Eur J Epidemiol 22:675–683

Glass TA (2000) Psychosocial intervention. In: Berkman LF, Kawachi I (eds) Social epidemiology. Oxford University Press, New York, pp 267–305

Glymour MM (2006) Natural experiments and instrumental variable analyses in social epidemiology. In: Oakes JM, Kaufman JS (eds) Methods in social epidemiology. Jossey-Bass, San Francisco, pp 423–445

Griffin J, Fuhrer R, Stansfeld SA, Marmot MG (2002) The importance of low control at work and home on depression and anxiety: do these effects vary by gender and social class? Soc Sci Med 54:738–798

Grossman AB (1991) Regulation of human pituitary responses to stress. In: Brown MB, Koob GF, Rivier C (eds) Stress: neurobiology and neuroendocrinology. Marcel Dekker, New York, pp 151–171

Hannan PJ (2006) Experimental social epidemiology: controlled community trials. In: Oakes JM, Kaufman JS (eds) Methods in social epidemiology. Jossey-Bass, San Francisco, pp 335–363

Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, Slagboom PE, Lumey LH (2008) Persistent epigenetic differences associated with prenatal exposure to famine in humans. Proc Natl Acad Sci USA 105:17046–17049

Hellhammer DH, Buchtal J, Gutberlet I, Kirschbaum C (1997) Social hierarchy and adrenocortical stress reactivity in men. Psychoneuroendocrinology 22:643–650

Hemingway H, Shipley MJ, Christie D, Marmot M (1998) Cardiothoracic ratio and relative heart volume as predictors of coronary heart disease mortality: the Whitehall study 25 year follow up. Eur Heart J 19:859–869

Hertzman C, Power C, Matthews S, Manor O (2001) Using an interactive framework of society and lifecourse to explain self-rated health in early adulthood. Soc Sci Med 53:1575–1585

Holtzman NA (2002) Genetics and social class. J Epidemiol Community Health 56:529–535

Howden-Chapman P, Matheson A, Crane J, Viggers H, Cunningham M, Blakely T, Cunningham C, Woodward A, Saville-Smith K, O'Dea D, Kennedy M, Baker M, Waipara N, Chapman R, Davie G (2007) Effect of insulating existing houses on health inequality: cluster randomised study in the community. Br Med J 334:460

Howlet T, Rees L, Besser G (1985) Cushing's syndrome. Clin Endocrinol Metab 14:911–945

Illsley R (1986) Occupational class, selection and the production of inequalities in health. Q J Soc Aff 2:151–165

Kaplan GA, Keil JE (1993) Socioeconomic factors and cardiovascular disease: a review of the literature. Circulation 88:1973–1998

Karasek R, Baker D, Marxer F, Ahlbom A, Theorell T (1981) Job decision latitude, job demands and cardiovascular disease: a prospective study of Swedish men. Am J Public Health 71:694–705

Kasl SV, Jones BA (2000) The impact of job loss and retirement on health. In: Berkman LA, Kawachi I (eds) Social epidemiology. Oxford University Press, New York, pp 118–136

Kaufman JS, Cooper RS (1999) Seeking causal explanations in social epidemiology. Am J Epidemiol 150:113–120

Kawachi I, Berkman LA (2000) Social cohesion, social capital and health. In: Berkman LF, Kawachi I (eds) Social epidemiology. Oxford University Press, New York, pp 174–190

Kawachi I, Kennedy B, Wilkinson R (1999a) Crime: social disorganization and relative deprivation. Soc Sci Med 48:731

Kawachi I, Kennedy BP, Glass R (1999b) Social capital and self-rated health: a contextual analysis. Am J Public Health 89:1187–1193

Kelly MP, Morgan A, Bonnefoy J, Butt J, Bergman V (2007) The social determinants of health: developing an evidence base for political action. Final report to World Health Organization Commission on the Social Determinants of Health from Measurement and Evidence Knowledge Network. 2007. WHO, Geneva www.who.int/social_determinants/resources /mekn_report_10oct07.pdf

Kemper TD (1993) Sociological models in the explanation of emotions. In: Lewis M, Haviland JM (eds) Handbook of emotions. The Guildford Press, New York, pp 41–52

Krieger N (2001a) A glossary for social epidemiology. J Epidemiol Community Health 55: 693–700

Krieger N (2001b) Commentary: society, biology and the logic of social epidemiology. Int J Epidemiol 30:44–46

Kristenson M, Orth-Gomer K, Kucinskiene Z, Bergdahl B, Calkauskas H, Balinkyiene I, Olsson AG (1998) Attenuated cortisol response to a standardised stress test in Lithuanian vs. Swedish men. Int J Behav Med 5:17–30

Kubzansky LD, Kawachi I (2000) Affective states and health. In: Berkman LF, Kawachi I (eds) Social epidemiology. Oxford University Press, New York, pp 213–241

Kuh D, Ben-Shlomo Y (1997) A life course approach to chronic disease epidemiology. Oxford University Press, Oxford

Kumari M, Badrick E, Chandola T, Adler NE, Epel E, Seeman T, Kirschbaum C, Marmot MG (2010) Measures of social position and cortisol secretion in an aging population: findings from the Whitehall II study. Psychosom Med 72:27–34

Lynch JW, Kaplan GA, Salonen JT (1997) Why do poor people behave badly? Variation in adult health behaviours and psychosocial characteristics by stages of the socioeconomic lifecourse. Soc Sci Med 44:809–819

Macintyre S, Ellaway A (2000) Ecological approaches: rediscovering the role of the physical and social environment. In: Berkman LF, Kawachi I (eds) Social epidemiology. Oxford University Press, New York, pp 332–348

Macintyre S, Maciver S, Sooman A (1993) Area, class and health: should we be focusing on places or people? J Soc Policy 22:213–234

Maes M, Meltzer H, Stevens W, Calabrese J, Cosyns P (1994) Immuendocrine aspects of major depression. Relationships between plasma interleukin-1 and soluble interleukin-2 receptor, prolactin and cortisol. Prog Neuropsychopharmacol Biol Psychiatry 18:717–730

Marmot MG, Rose G, Shipley M, Hamilton PJS (1978) Employment grade and coronary heart disease in British civil servants. J Epidemiol Community Health 32:244–249

Marmot MG, Shipley MJ (1996) Do socioeconomic differences in mortality persist after retirement? 25 year follow up of civil servants from the first Whitehall study. Br Med J 313:1177–1180

McGowan PO, Sasaki A, D'Alessio AC, Dymov S, Labonte B, Szyf M, Turecki G, Meaney MJ (2009) Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse. Nat Neurosci 12:342–348

McKeown T (1979) The role of medicine: dream, mirage or nemesis? Basil Blackwell, Oxford

Meaney MJ, Szyf M (2005) Environmental programming of stress responses through DNA methylation: life at the interface between a dynamic environment and a fixed genome. Dialogues Clin Neurosci 7:103–123

Menzhaghi F, Heinrichs S, Pich E, Weiss F, Koob G (1993) The role of limbic and hypothalamic corticotropin releasing factor in behavioural response to stress. Ann N Y Acad Sci 697:142–154

Mishra G, Nitsch D, Black S, De SB, Kuh D, Hardy R (2009) A structured approach to modelling the effects of binary exposure variables over the life course. Int J Epidemiol 38:528–537

Mroczek D, Kolarz C (1998) The effect of age on positive and negative affect: a developmental perspective on happiness. J Personal Soc Psychol 75:1333–1349

Muntaner C (2001) Social epidemiology: no way back. A response to Zielhus and Kiemeney. Int J Epidemiol 30:625–626

Navarro V, Shi L (2001) The political context of social inequalities in health. Soc Sci Med 52: 481–491

Oakes JM, Johnson PJ (2006) Propensity score matching for social epidemiology. In: Oakes JM, Kaufman JS (eds) Methods in social epidemiology. Jossey-Bass, San Francisco, pp 364–386

Pollitt RA, Rose KM, Kaufman JS (2005) Evaluating the evidence for models of life course socioeconomic factors and cardiovascular outcomes: a systematic review. BMC Public Health 5:7

Power C, Hertzman C (1997) Social and biological pathways linking early life and adult disease. Br Med Bull 53:210–221

Putnam R (2000) Bowling alone: the collapse and revival of American community. Simon and Schuster, New York

Rose D, Pevalin D (2002) The National Statistics Socio-economic Classification: unifying official and sociological approaches to the conceptualisation and measurement of social class. Soc Contemp 45/46:75–106

Rose G (1992) The strategy of preventive medicine. Oxford University Press, Oxford

Rothman K, Adami H, Trichopoulos D (1998) Should the mission of epidemiology include the eradication of poverty? Lancet 352:810–813

Sapolsky R, Krey L, McEwen B (1986) The neuroendocrinology of stress and aging: the glucocorticoid cascade hypothesis. Endocr Rev 3:301

Sapolsky RM (1993) Endocrinology alfresco: psychoendocrine studies of wild baboons. Recent Prog Hormone Res 48:437–468

Seidell JC, Bjorntorp P, Sjostrom L, Kvist H, Sannerstedt R (1990) Visceral fat accumulation in men is positively associated with insulin, glucose, and C-peptide levels, but negatively with testosterone levels. Metabolism 39:897–901

Selye H (1956) The stress of life. McGraw-Hill, New York

Siegrist J (1996) Adverse health effects of high-effort/low-reward conditions. J Occup Health Psychol 1:27–41

Slogett A, Joshi H (1994) Higher mortality in deprived areas: community or personal disadvantage? Br Med J 309:1470–1474

Snow J (1855) On the mode of communication of cholera, 2 edn. Churchill, London

Spiegel D, Bloom J, Kraemer H, Gottheil E (1989) Effect of psychosocial treatment on survival of patients with metastatic breast cancer. Lancet 14:888–891

Spielberger CD, Krasner SS (1988) The assessment of state and trait anxiety. In: Noyes R Jr, Roth M, Burrows GD (eds) Enological factors and associated disturbances. Elsevier Science Publishers B.V., Holland, pp 31–50

Stansfeld SA, Head J, Marmot MG (1998) Explaining social class differences in depression and well-being. Soc Psychiatry Psychiatr Epidemiol 33:1–9

Suomi SJ (1997) Early determinants of behaviour: evidence from primate studies. Br Med Bull 53:170–184

Susser M (1995) The tribulations of trials-interventions in communities. Am J Public Health 85:156–158

Syme SL, Berkman LF (1976) Social class, susceptibility, and sickness. Am J Epidemiol 104:1–8

Szreter S (1988) The importance of social intervention in Britain's mortality decline c.1850–1914: a re-interpretation of the role of public health. Soc Hist Med 1:1–37

The Marmot Review (2010) Fair society, healthy lives. Strategic review for health inequalities in England post 2010. The Marmot Review, London

Walberg P, McKee M, Shkolnikov V, Chenet L, Leon D (1998) Economic change, crime and mortality crisis in Russia: regional analysis. Br Med J 317:312–318

Waldron I, Hughes M, Brooks T (1996) Marriage protection and marriage selection – Prospective evidence for reciprocal effects of marital status and health. Soc Sci Med 43:113–123

Ware J, Gandek B (1998) Overview of the SF-36 health survey and the International Quality of Life (IQOLA) Project. J Clin Epidemiol 51:900–912

Weaver IC, Cervoni N, Champagne FA, D'Alessio AC, Sharma S, Seckl JR, Dymov S, Szyf M, Meaney MJ (2004) Epigenetic programming by maternal behavior. Nat Neurosci 7:847–854

West P (1991) Rethinking the health selection explanation for health inequalities. Soc Sci Med 32:373–384

Zielhus GA, Kiemeney LA (2001) Social epidemiology? No way. Int J Epidemiol 30:43–44

Occupational Epidemiology

42

Franco Merletti, Dario Mirabelli, and Lorenzo Richiardi

Contents

F. Merletti (✉) • D. Mirabelli • L. Richiardi
Unit of Cancer Epidemiology and Centre for Oncologic Prevention, Department of Medical Sciences, University of Turin, Turin, Italy

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_24,

42.1 Introduction

Occupational epidemiology has the same main goal as the broad field of epidemiology: to identify the causes of disease in a population in order to intervene to remove them. Occupational epidemiology is an exposure-oriented discipline; it is thus the systematic study of illnesses and injuries related to the workplace environment (Checkoway et al. 2004).

The first concern about occupational causes of disease may have been that of Hippocrates, who wrote about the lifestyle habits and environment of populations and patients. Nevertheless, it was the Italian physician Bernardino Ramazzini who recommended that doctors add questions about occupation to those recommended by Hippocrates, and it was Ramazzini who made the first systematic description of occupational diseases and their causes in his book *De Morbis Artificum* (Ramazzini 1964). His descriptions included different characteristics of skin ulceration in freshwater and sea fishermen, silicosis among stonemasons, ocular disorders among glassblowers, and neurological toxicity among tradesmen exposed to mercury. It is noteworthy that he not only described the diseases but was also deeply concerned about the ethics of harmful work practices and the need for preventive measures, such as good ventilation and protective clothing.

Classic historical reports, such as those about scurvy in sailors in 1753, scrotal cancer in chimney sweeps in 1775, respiratory cancers in underground metal miners in 1879, and bladder cancer in dye workers in 1895, are clear examples of the importance of reports of case series by clinicians and by the workers themselves (Carter 2000). New occupational hazards came to light incidentally even in the mid-1900s, when the methodological landmark of the historical cohort study was designed (Doll 1952, 1955; Case et al. 1954) and occupational epidemiology developed as a discipline. Indeed, Case and coauthors suspected that rubber workers would have an elevated risk for bladder cancer while conducting a study on the high incidence of bladder tumors among dye manufacturers (Doll 1975). While reviewing hospital records of bladder cancer patients in Birmingham, England, chosen as a control area because it did not have a dye industry, they noticed that many workers had been employed in a rubber factory. Subsequent investigation confirmed the association with rubber production and showed that it resulted from exposure to an antioxidant containing the carcinogen 2-naphthylamine (Case and Hosker 1954; Coggon 2000).

Occupational epidemiology has contributed to the development of both study designs (such as the historical cohort study) and analytical methods that are now part of the broader field of epidemiology and of other exposure-oriented disciplines. For instance, quantitative and qualitative methods for assessing exposure, such as job-exposure matrices and job-specific questionnaire modules for assessment by experts, were developed by occupational epidemiologists and industrial hygienists. They have now been adapted and used in other disciplines, such as nutritional and environmental epidemiology, and are central to ensuring the validity and informativeness of epidemiological research in general.

Prevention is the final goal of all epidemiological research and findings. Occupational exposure was one of the first causes to be identified of diseases such as cancer and pulmonary illness, and epidemiological study of such exposures often led to the identification of specific causal agents. Occupational hazards are known causes of disease that are amenable to regulatory control and thus especially suitable for prevention. This is in contrast to aspects of lifestyle, such as smoking and dietary habits, for which control requires modification of cultural and personal behavior patterns. Free choice may contribute to some diseases attributable to environmental causes; for instance, the large majority of cases of lung cancer are attributable to tobacco smoking and can be prevented by avoiding the habit. The reason for interest in preventing occupational hazards is more subtle: As personal choice plays little or no role in occupational exposure, the protection of workers warrants special attention. Furthermore, while industrial effluents and products might cause illness in the general population, exposed workers are likely to be the first and most severely affected. Prevention at the level of the working environment will by the same token result in prevention in the general population.

This chapter will address issues in study designs and epidemiological methods as applied in the specific field of occupational epidemiology. They will include dose-response analysis, healthy worker effect, and exposure assessment. Finally, how occupational epidemiology can help to evaluate the need and effectiveness of primary prevention interventions and policies will be described using the example of occupational cancer.

42.2 Study Designs

Classic epidemiological studies, such as cross-sectional (see chapter ►Descriptive Studies of this handbook), case-control (see chapter ►Case-Control Studies of this handbook), and cohort studies (see chapter ►Cohort Studies of this handbook), are commonly carried out in occupational settings. The principles of study design and data analysis are derived from general epidemiological methods; however, some specific aspects are worth addressing.

42.2.1 Cross-Sectional Studies

Cross-sectional studies are generally used to investigate symptoms, sub-clinical outcomes, and physiological functions that are (or are suspected to be) related with an exposure of interest. Examples are wheezing and other respiratory symptoms in relation to exposure to different types of metal-working fluids (Greaves et al. 1997), urinary mutagenicity and DNA adducts in peripheral blood mononuclear cells and urothelial cells in relation with urinary 1-hydroxypyrene levels (Peters et al. 2008), and forced expiratory volume in one second (FEV_1) and diesel exhaust exposure (US Environmental Protection Agency 2002). Cross-sectional studies are also useful in the study of potentially reversible and non-fatal diseases, such as musculoskeletal disorders. For all of such conditions, longitudinal studies are either inappropriate or not practically feasible, and cross-sectional studies are the best study design applicable. It must be born in mind, however, that they measure the prevalence of the outcome of interest, or assess a dose-response relationship, in groups that are often an opportunistic, rather than a representative, random sample of the ideal target population (e.g., all metal or all rubber workers), being made up of the individuals still active at the time of the survey and willing to participate. Therefore, associations between exposure and disease are difficult to interpret, as they could depend either on an increased incidence or on a longer duration of disease among a subgroup of cases. For this reason and for problems of reverse causality arising from measuring exposures and diseases at the same time, which prevents the time sequence between exposure and outcome to be clearly established, the causal nature of an association can be weakly addressed using a cross-sectional approach.

Cross-sectional studies allow the researchers to examine on an ad hoc basis the target population, or a sample of it, and by doing so to study health outcomes that cannot be investigated in longitudinal studies of mortality or incidence. At the same time, they are vulnerable to the effect of non-response, particularly when they are carried out with the main aim of estimating the prevalence of diseases or their symptoms. Diseased workers may participate in the study differently from those who are not diseased, and their willingness to participate may depend on their exposure status. Occupational studies of fertility and sperm quality are an example of studies in which non-response is a critical problem. Since the observation of the toxic effects of 2,3-dibromo-3-chloropropane on testicular germ cells (Potashnik et al. 1978), the fertility of exposed male workers has been investigated in several studies. In one study, groups of traditional and organic farmers were selected randomly from the database of the Danish Ministry of Agriculture in 1995–1996 and invited to participate in a study on semen quality, including total sperm count, sperm concentration, other indices, and serum concentrations of sex hormones (Larsen et al. 1999). A questionnaire eliciting information on previous exposure to pesticides was posted to 1,124 farmers, of whom 86% answered and 256 provided semen samples. This low participation proportion was not unexpected, as the examination required by the study was somewhat demanding.

A further limitation of cross-sectional studies, which is specific to occupational epidemiology, is that only active workers are usually investigated, because the study base is defined as workers employed in a specific industry or exposed to a specific occupational factor. It follows that workers who have terminated their employment cannot be included in the study.

Let us consider the example of the cross-sectional studies on the health effects of exposure to diesel fumes (US Environmental Protection Agency 2002). Acute respiratory effects were investigated in several studies by measuring FEV_1 and other indicators of pulmonary function twice, at the beginning and at the end of a work shift, in workers employed in mines and garages. Chronic respiratory effects were studied through a single survey and a medical examination in workers with different levels of cumulative occupational exposure to diesel exhausts. Individuals who are susceptible to diesel exhaust exposure tend to move from jobs with a high level of exposure. Therefore, a cross-sectional study on the acute effects of exposure is presumably carried out among a selected group of workers, resulting in a possible underestimate of the effects. Regarding chronic effects, which are manifest a long period after the exposure has occurred, there is an underestimate of the association between exposure and disease, if the termination of employment is determined by the disease or its early symptoms.

Notwithstanding these limitations, cross-sectional studies have been successfully carried out on the association between certain exposures and long-term effects, taking advantage of the fact that this study design offers the opportunity to investigate exposures thoroughly. A group of French talc millers received a respiratory health examination in 1989, including a standardized questionnaire on respiratory symptoms, lung function testing, and a chest radiograph (Wild et al. 1995). Systematic measurements of dust concentrations had been made in 1986 and 1989, and, based on the results of the measurement campaigns and of an accurate retrospective reconstruction of past working conditions, a quantitative job-exposure matrix had been developed. The cumulative exposure of all workers included in the respiratory health survey was estimated, by applying the matrix to their work histories. A dose-response relationship between cumulative dose and decline in lung function, prevalence of dispnoea, and chest X-ray opacities could be shown.

The initial survey of a cross-sectional study may become the first phase of a longitudinal study. In the investigation of French talc millers, the same group of workers included in the 1989 respiratory health survey received a second chest X-ray in 1993. An increase was observed in the number of workers with profusion of opacities >1/0, but the increase was not associated with cumulative exposure to talc (Wild et al. 1995).

42.2.2 Cohort Studies

To investigate long-term effects, the cohort study is a valid, but sometimes expensive and time-consuming, design. Nevertheless, the availability of employment records and trade union registries often permits straightforward identification of past occupational cohorts. It is therefore not surprising that historical cohort studies have long been the method of choice in occupational epidemiology, and they have contributed significantly to the identification of occupational hazards.

Researchers usually identify a factory in which the exposure of interest occurs – to specific chemicals and substances or specific working conditions and job tasks – and select the members of the cohort from registries available at the factory. Alternatively, a study population can be identified from similar departments in different factories. Thus, when a single facility does not provide a sufficient number of workers or the time of follow-up is not long enough, a collaborative study can be conducted in similar factories in several centers.

The cohort study of workers employed in the man-made vitreous fiber (MMVF) industry in Europe, coordinated by the International Agency for Research on

Cancer (IARC) (Boffetta et al. 1997, 1999; Sali et al. 1999), is an example of such collaboration. The cohort was assembled in 1977 and consisted of approximately 22,000 workers who had ever been employed in 13 factories producing at least one of three types of MMVF, namely, glass wool, continuous filaments of glass fiber, and rock or slag wool, at any time between the year of starting production of MMVF and 1977. The follow-up ended between 1990 and 1995 in different factories, depending on subsequent updating.

Exposure was assessed on the basis of individual work histories, obtained from employment registries in 1977. It was known that important technological changes had taken place in the production of MMVF over the study period, so that the period of MMVF production was divided into three "technological phases": early, intermediate, and late. As the ambient levels of exposure to MMVF were estimated to have decreased with evolving production processes, the year in which each of the three phases began in each factory was assessed. Information on possible concomitant exposure to other agents, such as asbestos and bitumen, was also obtained for each factory. The researchers thus knew the duration of employment for each worker, by factory, technological phase, and job task.

National mortality rates were used to calculate standardized mortality ratios (*SMR*s) (cf. chapters ▸Rates, Risks, Measures of Association and Impact or ▸Descriptive Studies of this handbook) for neoplastic and non-neoplastic causes of death, and cancer-specific standardized incidence ratios (*SIR*s) were estimated for the subcohorts in countries where cancer incidence rates are available from cancer registries. The effect of duration of employment was estimated in internal comparisons within the cohort, the reference group including workers employed for less than 5 years. Data were also analyzed according to type of MMVF, job task, technological phase, and time since first employment. Data on workers who had been employed for less than 1 year were analyzed separately, as short-term workers might be high-risk individuals with particular lifestyles or occupational exposure to agents other than MMVF (see also Sect. 42.4.2).

In general, MMVF production workers did not have an excess risk of mortality or cancer incidence, although a small excess risk for lung cancer was found among rock- and slag-wool workers, and increased mortality from heart diseases and non-malignant renal diseases was suggested. It is important to note that no information on lifestyle factors was available, which is a limitation of almost all historical cohort studies.

In countries where good, computerized population registries with a long history of registration exist, large occupational cohort studies can be carried out by linkage of information on occupational status from censuses with individual data on, for example, mortality, cancer incidence, and hospital discharges. The strength of record-linkage studies is the very large sample size. An occupational record-linkage study of cancer incidence was conducted in the Nordic countries among persons aged 25–64 years who were listed in the 1970 censuses (Andersen et al. 1999). Overall, about ten million persons were included in the study, and more than 500,000 incident cases of cancer occurred during the follow-up period, which ended between 1987 and 1990, depending on the country. Occupational exposure was evaluated for 54 occupational groups. Many cancer-specific associations were estimated, and they cannot be discussed here; however, the general finding was that risk of cancer is associated with occupation.

This record-linkage study shows clearly that cohort studies can provide risk estimates for many outcomes and some of the findings might be unexpected. For instance, in a historical cohort study of 8,226 workers employed in an aircraft manufacturing factory in northern Italy between 1954 and 1981, an unexpected excess of melanomas was found (6 observed, 1.02 expected cases) (Costa et al. 1989). When an unexpected association is found, the characteristics of the cases, including age, sex, period of employment, factory, and job task, should be explored carefully in order to identify any clusters of jobs or operations that suggest a common exposure. In the example of melanoma, the characteristics of the six cases were described in detail, but no cluster could be identified.

There are two major limitations to the use of data from existing records rather than from ad hoc questionnaires and environmental or biological measurements: lack of detailed information on exposure and lack of information on possible relevant confounders.

With regard to exposure, maximum cooperation between researchers and management, trade unions, occupational physicians, and industrial hygienists is crucial to obtain information on the nature of both industrial processes and working environments. Basic information on the exposure of each worker should include the starting and ending dates of employment at the factory. Unfortunately, important information, such as the job task of each worker and changes in industrial processes over time, is often missing. Even when the job task is recorded, one would like to evaluate also the variability of exposure levels among workers carrying out the same job. The general lack of information may reduce the quality of the data on exposure, whatever approach is used to assess exposure, and finally bias the results of the study because of misclassification. Although it is theoretically possible to measure factory-specific levels of exposure at the time a study is conceived, strong assumptions should hold for a reliable imputation of past exposures.

In some studies, plant-specific ambient measurements had been recorded over time and were available for assessing exposure. A historical cohort of more than 74,000 workers employed between 1972 and 1987 in 672 factories in jobs that entailed exposure to benzene was assembled in China (Hayes et al. 1997). The cohort was followed-up for death from all causes and for incidence of hematological tumors, with an analogous cohort of approximately 36,000 unexposed workers for comparison. For the purposes of assessing exposure, information on the factory and department of employment and on the starting and ending dates of each job was obtained, for each worker, from employment records available at the factories (Dosemeci et al. 1994a). Moreover, information on production activities and changes in processes over time was obtained at each factory and for each job type. Importantly, the results of all past air monitoring (more than 8,400 measurements) for benzene and other solvents were also obtained. Therefore, whenever possible, the exposure level was assigned to each worker on the basis of monitoring results either for specific combinations of job task, department, and calendar period or for adjacent calendar periods or similar job tasks.

Detailed information on exposure and confounders can obviously be obtained in concurrent cohort studies, which can be efficiently carried out when the induction period between exposure and disease is short. If the cohort is followed-up prospectively, temporal variations in exposure can be ascertained either at individual level, from questionnaires, personal dosimetry data, or use of biomarkers of exposure, or at aggregate level, from environmental measurements and monitoring of changes in industrial processes.

42.2.3 Case-Control Studies

Studies in Industrial Populations Nested case-control studies (see chapter ▸Modern Epidemiological Study Designs of this handbook) might solve some of the limitations inherent in the cohort design. As a nested case-control study covers fewer persons than a cohort study, the nested approach is efficient when the exposure assessment is not straightforward, as, for instance, when it is based on experts' judgment. The nested approach is also more efficient when worker-specific levels of exposure are estimated from biological samples or by direct interview with the workers or their next-of-kin. Measurement of biomarkers can result in accurate assessments of current exposure, but assumptions must be made about past exposure. Conversely, interviews allow detailed reconstructions of working histories and provide information on possible confounders. Information on actual exposure levels may nevertheless be rather imprecise, and the subjects are difficult to trace, especially when the follow-up period is long.

The historical cohort study of workers employed in MMVF production coordinated by the IARC, described above, includes a clear example of a nested case-control design (Boffetta et al. 1997). The analyses of the cohort revealed a small excess risk for lung cancer among rock- and slag-wool production workers, but no information was available on possible confounders; furthermore, occupational histories were available up to 1977 only and were limited to the information in the employment registries. The researchers therefore conducted a case-control study of 196 cases of lung cancer, and 1,715 matched controls nested in the cohort (Kjaerheim et al. 2002). The index subjects or their next-of-kin were traced and interviewed to obtain information on lifetime smoking habits, residential history, and lifetime occupational history, both within and outside the MMVF industry. As anticipated by the study design, the proportion of completed interviews with the selected subjects was low: 68% for cases and 35% for controls. Two industrial hygienists evaluated the individual occupational histories for exposure to each of several occupational agents known or suspected to be associated with lung cancer. Moreover, an expert panel was formed to evaluate individual cumulative exposure to MMVF on the basis of the new information obtained at the interview. The smoking-adjusted estimates and the analyses by quartiles of cumulative level of exposure in the nested study did not support an association between exposure to rock- or slag-wool and lung cancer risk.

Quite often, a nested case-control design increases the efficiency of the computerization, cleaning, and handling of data, even though information on exposure is available. Grayson (1996), for example, conducted a case-control study on brain cancer nested in a cohort of approximately 880,000 US Air Force members to evaluate the effect of occupational exposure to electromagnetic fields. The workers had to have been employed between 1970 and 1989. At the end of the follow-up period, 230 incident cases of brain cancer were found, and four controls for each case were randomly selected among cohort members. Information on past exposure to electromagnetic fields was obtained from several sources, including employment records, records of events exceeding existing limits, and some personal dosimetry data. The final analysis was based on 1,150 persons instead of more than 800,000 in the original cohort.

Studies in the General Population Population- or hospital-based case-control studies have frequently been used to investigate the health effects of occupational exposures. In the early 1980s, a multicenter case-control study was carried out to investigate the associations between laryngeal and hypopharyngeal cancer and smoking, alcohol, dietary habits, and occupational factors (Tuyns et al. 1988). The study, coordinated by IARC, was population-based and included six centers in northern Italy, France, Spain, and Switzerland. Information on occupational history and lifestyle factors was obtained by face-to-face interviews with cases and controls. Specifically, each person was asked to report all jobs held for at least 1 year since 1945, specifying their starting and ending years, a short description of specific tasks, the name of the company, the company's activity, and the specific products of the department in which the interviewed person had worked. The occupational histories of 1,010 interviewed cases and 2,176 interviewed controls were coded, without knowledge of case or control status, according to standard international classifications of occupations and industries. Then, smoking- and alcohol-adjusted odds ratios for occupational factors were obtained by two approaches. First, an exploratory analysis was carried out on 156 occupations and 70 industrial activities in which at least nine individuals had been ever employed (Boffetta et al. 2003). Second, a working group created a job-exposure matrix (JEM) to categorize each combination of job and activity in terms of levels of probability, intensity, and frequency of exposure to 16 occupational agents for which there was some a priori evidence of an association with laryngeal cancer risk (Berrino et al. 2003). The agents investigated included asbestos, solvents, formaldehyde, and polycyclic aromatic hydrocarbons. The JEM was used and evaluated in ad hoc studies (Merletti et al. 1991; Ahrens et al. 1993; Luce et al. 1993; Orlowski et al. 1993; Stengel et al. 1993; Stucker et al. 1993). An account of its validation, based on a comparison between the results of the JEM and the experts' evaluation of the jobs as described in the questionnaires, is given in Table 42.1. Generally, the specificity and sensitivity of the JEM was agent-specific. The first analytical approach, based on job titles and industrial activities, provided risk estimates for several occupations, an advantage

Table 42.1 Validation of the job exposure matrix (JEM) of the IARC case-control study on laryngeal and hypopharyngeal cancer: percentage of jobs not entailing an exposure to specific agents according to an expert's assessment compared with the results from the JEM

	JEM categories of intensity/probability of exposure[b]							
Agent[a]	No. of job periods	1	2	3a	3b	3c	4	5
Asbestos	3220	96	83	79	73		68[c]	
Solvents (1)	2712	96	92	89	70	47	58	16
Solvents (2)	929	87	83	62	67	35	37	9
Formaldehyde	884	75	90	59	47	50	29	–[d]
Wood dust	863	95	–[d]	50	50	–[d]	8	0
PAH	2571	98	68	88	85	98	74	39

[a] Agents were evaluated in the following studies: asbestos, Orlowski et al. (1993); solvents (1 = bladder cancer study, 2 = glomerulonephritis study), Stengel et al. (1993); formaldehyde and wood dust, Luce et al. (1993); *PAH* polycyclic aromatic (hydrocarbons), Stucker et al. (1993)

[b] Categories: 1. Job-related exposure is not higher than for the general population; 2. job entails or may entail a cumulative exposure slightly higher than for the general population; 3. job may entail exposure definitely higher than for the general population, but the coded information is not sufficient to discriminate between exposed and not exposed workers (3a: few workers are thought to be exposed, and 3b: some workers are thought to be exposed, 3c: the majority of workers are thought to be exposed); 4. job entails exposure to the specific agent at definitely higher level than the general population; 5. job entails exposure to the specific agent, and there are instances in which the exposure is known to be particularly high

[c] Categories 3c, 4, 5 were considered jointly

[d] Category with no jobs according to the JEM

facilitated by the heterogeneity of the study subjects' working histories due to the multicenter design. Conversely, the second approach directly tested etiological hypotheses. In both instances, the case-control design made it possible to control for the confounding effects of smoking, alcohol drinking, and diet.

Mortality Odds Ratio Studies Mortality odds ratio studies have a case-control design and are a valid alternative to proportionate mortality studies, which have been widely used in occupational epidemiology (Miettinen and Wang 1981; Boyd et al. 1970). In proportionate mortality studies, the frequency of death for the diseases under study among exposed workers is compared with the corresponding figure calculated for a reference population (proportionate mortality ratio, *PMR*). *PMR*s are limited by the fact that they must add up to unity; therefore, elevated *PMR*s for some diseases are, by definition, counterbalanced by decreased *PMR*s for other diseases. Moreover, *PMR*s are biased if ascertainment of deaths is incomplete in a different proportion among exposed than unexposed subjects. These drawbacks are overcome in mortality odds ratio studies where the case-control approach is applied. In such studies, the cases comprise deaths from the specific cause of interest, both exposed and unexposed, while the controls are other deaths, selected on the basis of a presumed lack of association with the exposure. The principle of selecting the control causes of death for inclusion in the study is therefore the same as selecting a control series for any case-control study: Controls are selected independently of exposure and with the aim of representing the proportion of exposure in the study base (Rothman et al. 2008). In practice, however, avoiding bias may prove difficult, as many causes of death, even when unrelated with the exposure under study, are associated indirectly with occupational exposures via the socioeconomic status.

Concluding Remarks Case-control studies can be used efficiently to investigate long-term outcomes represented by rare diseases, or by diseases caused by widespread occupational exposures, which cannot be localized to a specific industry. This study design also permits the researcher to focus on minorities and on subgroups of the population that have often been poorly investigated. For example, at the first international conference on occupational cancer in women, in 1993, it was recognized that most of the information on occupational hazards had been obtained from studies on men: A survey showed that less than 10% of published epidemiological studies included and reported detailed results on women (Zahm et al. 1994). Although this picture has changed, efforts to study the effects of occupational exposures on women are still needed (Zahm and Blair 2003). A case-control approach is often used in occupational epidemiology for exploratory studies. As in the study on laryngeal and hypopharyngeal cancer described above, an occupational history may be classified by several groups of job titles and industrial activities. As multiple comparisons are made, a Bayesian approach with semi-Bayes or empirical-Bayes adjustments might help to decrease the impact of false-positive results (Greenland and Poole 1994; Corbin et al. 2008). For a formal explanation and practical examples of Bayesian approaches in occupational epidemiology, see Greenland and Poole (1994) and Steenland et al. (2000a).

42.3 Exposure Assessment

Exposure assessment (see chapter ▸Exposure Assessment of this handbook), a critical step in any epidemiological study, is central in occupational epidemiology. The most recent developments in the design of both cohort and case-control studies of work-related diseases rely on identification of exposure to specific agents, such as chemicals, rather than on the use of surrogates of exposures, such as being employed in a given industrial activity or holding a certain job. Furthermore, an attempt is often made to compute some measure of dose, such as cumulative exposure or average exposure, which in turn requires estimation of the level (intensity) of exposure and its variation over time.

42.3.1 Exposure Assessment in Industry-Based Studies

Two general strategies, statistical and deterministic modeling (Kauppinen et al. 1994), are available to assess exposure on the basis of the primary information collected in a study. Such information usually does not include satisfactory measures of exposure to the agent(s) of interest and is often limited to a description of the work setting, the operations performed by workers, and the materials they handled. When suitable data are available, however, on the concentration of the agent of interest in the workplace, exposure assessment will be based on a stochastic (statistical) approach, in which a model is fitted to the results of past industrial hygiene measurements by plant, job title, and work area; missing data will be estimated from the model (Kauppinen et al. 1994). When statistical modeling is applied to industry-based studies, such as cohort, cross-sectional, and nested case-control studies, workers are classified into "homogeneous" groups on the basis of combinations of plant, work area, job title, and period. The available industrial hygiene measurement series are broken down into the same groups (de Vocht et al. 2008; Lavoué et al. 2007). The main limitations of the statistical approach are that (1) trends in exposure over time are often unknown, either because measurements were not made for previous processes and working conditions or because of difficulties in the interpretation of historical measurements, and (2) the interindividual variation of exposure within a homogeneous group can be wider than that between groups, which may make the correct identification of homogeneous groups difficult (Kromhout et al. 1993). Furthermore, it must be born in mind that, when only limited data exist, they will also be distributed unevenly by job. There is, thus, the danger of overestimating exposures due to oversampling worst case scenarios, especially when measurements were initiated upon complaints of workers, for example, in response to spills. Conversely, measurements done by the companies or measurements made to assess compliance with exposure limits may lead to underrepresentation of the highest exposures.

In historical cohort studies, the availability and quality of industrial hygiene data are often different for the various settings included in the work histories of the study subjects; good data may exist for some periods and not for others. In these circumstances, the maximum achievable goal is a semiquantitative approach in which jobs are compared according to materials handled and tasks performed. The jobs are then ordered in terms of assessed exposure, which is placed onto a semiquantitative scale (e.g., high, intermediate, low).

Because comprehensive data on exposure are rarely available, less accurate methods have thus to be used. If the factors that determine the level of exposure can be identified, they can be used to construct a deterministic model (Kauppinen et al. 1994). In deterministic modeling in industry-based studies, the most significant factors that affect exposure intensity, such as type of plant and machinery, presence of local exhausts, and workers proximity to sources, are identified. Their relative importance is then assessed, either on the basis of historical industrial hygiene data – even if such data are insufficient for complete modeling – or, in their absence, through the theoretical evaluation of how different tasks, operations, and procedures could have affected exposure. In rare circumstances, it has been possible to improve the assessment by reproducing experimentally some past working conditions and to estimate the concentrations of the agents of interest by applying modern measurement techniques to such reconstructed working places and procedures. It has been shown that, by combining such approaches, complex industry-specific exposure matrices can be built on the basis of detailed knowledge of plant-, job-, and time-specific factors (Kauppinen and Partanen 1988). Semiquantitative exposure levels can then be established for each study subject applying the matrix to the information on the jobs they held and the tasks they performed.

The main limitations of the deterministic approach are that (1) the relative importance of the various determinants may prove difficult to assess, and agreement among experts may be poor, and (2) the quality of information on the determinants may be highly variable across study subjects; for some, the tasks involved in their job might have been recorded, while for others barely the job title or the department of assignment is known.

In some recent multicenter or pooled industry-based studies, considerable advances have been made in exposure assessment, by combining the statistical and deterministic approaches. Quantitative industry-specific exposure matrices were thus built and applied to the workers' individual job histories (Burstyn et al. 2000, 2003; t'Mannetje et al. 2002; Harber et al. 2003). Alternatively, Bayesian decision analysis may be applied: Experts will create a "prior" probability distribution of exposure, based on their knowledge of processes, jobs, etc., which will be combined with available measurements to obtain "posterior" distributions (Ramachandran 2001). Current standards of practice imply that when good industrial hygiene data are available, at least for some historical periods, and the relative influences of changes in plant, process, and activity can be evaluated, exposure will be assessed quantitatively and extrapolated to periods or plants for which no original quantitative information was available. With this method, quantitative data on a given job, in a given industrial activity, and during a given period provide a baseline estimate of both the average exposure and its variation. Known differences in the presence and characteristics of determinants provide multiplicative weighting factors to be applied to the baseline estimate. Few validation studies of industry-specific exposure matrices are, however, available (Dosemeci et al. 1994b, 1996).

42.3.2 Exposure Assessment in Population-Based Studies

In population- and hospital-based case-control studies, statistical modeling has been used to set up job-exposure matrices (JEM) (Hoar et al. 1980; Macaluso et al. 1983). A JEM can be defined as a cross-classification of a list of job titles with a list of agents to which the workers performing the jobs might be exposed (Kauppinen and Partanen 1988). Deterministic modeling has been used in the interpretation and assessment of job histories by industrial hygiene experts when occupational questionnaires including job-specific modules (JSM) were developed to obtain the detailed information necessary for the experts' judgment (Siemiatycki et al. 1981; Macaluso et al. 1983; Ahrens et al. 1993). Researchers at the US National Cancer Institute showed that a deterministic approach can be used not only in expert- and JSM-based assessment but also to create and use more detailed and improved JEMs that might allow semiquantitative or even quantitative exposure assessments (Dosemeci et al. 1990a). The same group suggested that the JEM-based assessment strategy should be abandoned in favor of the JSM-based expert assessment

(Stewart et al. 1996). Use of JEMs has been reported to result in loss of both sensitivity and specificity in exposure assessment, in comparison with the use of a JSM-based individual assessment (Rybicki et al. 1997). Simulation studies suggested that use of JEMs may lead to loss of precision in odds ratio estimates, whereas expert-based assessment resulted in relatively low levels of misclassification (Bouyer et al. 1995).

Although it is somewhat difficult to assess the validity of expert-based exposure assessment in the field, some studies suggest that the agreement within and between experts might be satisfactory when experienced teams of raters are available (Siemiatycki et al. 1997; Fritschi et al. 2003). Two studies addressed the issue of expert-based exposure assessment validation by means of an objective index of past exposure to asbestos.

The first study (Pairon et al. 1994) comprised 131 cases of mesothelioma. The probability, level, and frequency of exposure were assessed by using qualitative ordinal classifications of the job in which each person had maximum exposure. Combinations of assessed probability, level, and frequency were summarized in four classes: (1) unexposed or possibly exposed, (2) probable or definite exposure at low level, (3) probable or definite exposure at levels higher than low, with sporadic frequency, and (4) probable or definite exposure at levels higher than low, with more than sporadic frequency. No attempt to build up a cumulative dose index was made. A limited correlation between the exposure assessment and objective indices of exposure to asbestos was observed, particularly with counts of asbestos bodies per gram of dried tissue. This study suffered from some shortcomings. Intensity and frequency were not used to compute a combined dose estimate, which prevented the calculation of a cumulative dose index. Frequency was used to discriminate between the third and the fourth summary class, but variations in frequency might actually be less important than those in intensity to determine the average exposure level. Only the highest exposure job was used in the assessment, so that other possible exposures have not been taken into account. The sensitivity of objective indices of asbestos exposure in mesothelioma cases may be low.

In the second study (Takahashi et al. 1994), 42 cancer cases for whom necropsy material was available were assessed for exposure from a JSM-based questionnaire and by analysis of lung tissue fibers. A good correlation was found between the JSM-based exposure assessment and asbestos fiber counts, although some cases were found to have had exposure but had no asbestos in the lung. The main shortcomings of this study are its rather limited dimension and a potential necropsy selection bias; the heterogeneous nature of the cases as to their cancer site makes it difficult to extrapolate its results to a mesothelioma series.

Expert-based assessment with deterministic modeling in the hands of experienced raters has resulted in quantitative assessments in some population- and hospital-based case-control studies (Iwatsubo et al. 1998; Brüske-Hohlfeld et al. 2000; Rödelsperger et al. 2001; Pohlabeln et al. 2002).

Recently, nationwide databases of past measurements have been created by public institutions involved in health and safety at work, like Colchic at INRS (France), MEGA at Berufsgenossenschaften (Germany), or the National Exposure Database at the Health and Safety Executive (UK). The availability of large amounts of industrial hygiene data, going back in time to the 1970s and 1980s and earlier, is stimulating the development of quantitative JEMs (Kauppinen et al. 2009; Preller et al. 2010; Peters et al. 2012).

42.3.3 Consequences of Errors in Exposure Assessment

The consequences of errors in exposure assessment are discussed extensively in the chapters on exposure assessment (chapter ▸Exposure Assessment) and measurement error (chapter ▸Measurement Error) in this handbook. When exposure is measured as a continuous variable at the individual level, random, non-differential errors in assessment, such as those deriving from errors in measurement, generally lead to attenuation of the exposure-disease association and diminish the goodness of fit of regression models (Armstrong 1998). When measurements are constrained by a lower limit, such as a detection limit, however, inflation of the exposure-response association can occur under certain circumstances (Richardson and Ciampi 2003). When exposure is measured as a continuous variable, but at the group level, a rather different situation occurs: The exposure level is the average for a sample of individuals in the group, and all individuals are assigned this average exposure. This leads to what is referred to as "Berkson error." In a "classical" error, an individual is assigned a measured exposure, affected by random variability. In Berkson error, the group average exposure is affected by a considerably smaller random error, but the actual exposure of each individual in the group will be different from the group average. Berkson error usually entails small, if any, bias of the exposure-response association (Armstrong 1998), even if in certain circumstances a substantial over- or underestimation of the quantitative relationship may occur (Steenland et al. 2000b).

In many, probably most, study designs, exposure is scaled as a discrete variable on a dichotomous or a polytomous scale. When the exposure variable is dichotomized, non-differential misclassification will always bias the effect measure toward the null; however, when the exposure variable is polytomous, non-differential misclassification will bias the effect measure toward the null only if misclassification occurs between adjacent exposure categories. When it involves non-adjacent categories, bias away from the null may also occur (Armstrong 1998; Dosemeci et al. 1990b).

Quantitatively, misclassification is a function of (a) the sensitivity of the assessment method, that is, the proportion of all truly exposed subjects correctly classified as exposed and (b) its specificity, that is, the proportion of all truly unexposed subjects correctly classified as unexposed. The relative importance of sensitivity and specificity in overall misclassification bias depends on exposure prevalence: When exposures are rare, like most occupational exposures in population- and hospital-based case-control studies, even small losses in specificity may strongly bias the relative risk estimate toward the null. Such effect is clearly depicted in Fig. 42.1, where a true relative risk of 4.0, sensitivity in exposure assessment of 0.9, and a range of commonly found exposure prevalences (from 0.001 up to 0.2) are assumed. The estimated relative risk is plotted against different specificities in exposure assessment.

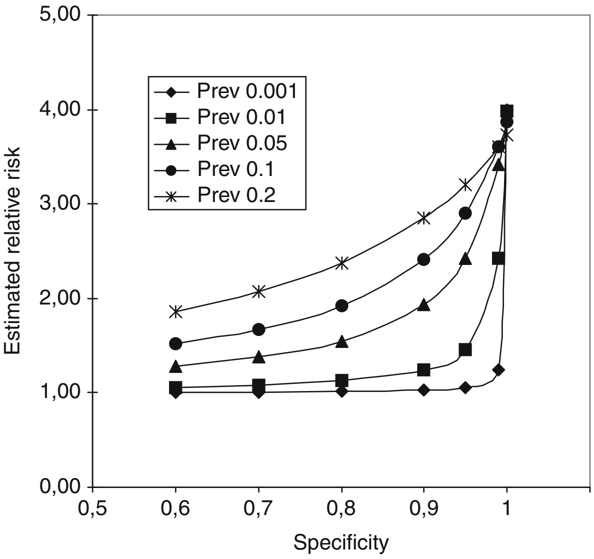


Fig. 42.1 Bias toward the null of estimated relative risk due to loss in specificity in exposure assessment. True relative risk = 4 and exposure assessment sensitivity 0.9 are assumed. Estimated relative risk is plotted against different levels of specificity of exposure assessment, for exposure with prevalences (Prev) ranging from 0.001 to 0.2 (Modified from Ahrens 1999)

This consideration does not of course imply that sensitivity is not important: When exposures are rare, low sensitivity in exposure assessment causes loss in power and requires substantial increases in sample size to compensate for it.

42.4 Special Issues in Occupational Epidemiology

42.4.1 Confounding

A general discussion of confounding is given in chapter ▸Confounding and Interaction of this handbook). Information on several known possible confounders and on other occupational and non-occupational exposures of interest is almost always lacking for historical occupational cohorts. Methods to deal with confounding in historical cohorts include use of internal comparison groups, with general characteristics assumed to be similar to those of the exposed subjects, and use of available statistics on the distribution of confounders in the population from which the cohort originated. The first approach is commonplace in occupational epidemiology, although it is seldom possible to verify whether the comparison group has the assumed characteristics. Internal comparisons have the advantage of controlling part of the bias introduced by the "healthy worker effect," that is discussed below. In a historical cohort study conducted in the Nordic countries to investigate the risks for cancer among airplane pilots (Pukkala et al. 2002), national cancer registry data were used to calculate standardized incidence ratios (*SIR*s). Since airplane pilots belong to a higher social class than the general population, however, the *SIR*s were possibly biased. In this study, cosmic radiation was the main exposure of interest; therefore, a cumulative dose of radiation experienced by each member of the cohort was calculated. This made it possible to check the main findings from the external comparison by analyzing the effect of the exposure, using pilots with the lowest cumulative dose as the reference group. Such a comparison is unlikely to be confounded by social class.

The second approach, the use of available population statistics, was applied in the Norwegian part of the occupational record-linkage study in the Nordic countries, described in Sect. 42.2.2, with occupation-, sex-, and birth-cohort-specific information on smoking habits in the population obtained from external surveys (Haldorsen et al. 2004). The Norwegian study evaluated 42 occupational groups for risk of lung cancer in comparison with twelve other occupational groups assumed to be without exposure to occupational lung carcinogens. The magnitude of the associations of proportion of current and former smokers and amount of cigarettes smoked with lung cancer risk was estimated among the twelve reference groups, using data from the external surveys. Then, the smoking-adjusted SIRs for the 42 occupational groups were calculated and compared with the non-adjusted estimates. Limitations of this approach are that (1) the quality of the information on the confounder depends on the quality of the surveys and (2) the magnitude of the association between confounder and disease is not directly estimated at an individual level. The magnitude of the association can be either obtained from the best available studies or from ad hoc studies conducted in the same population, or estimated, as in the Norwegian study, using aggregate data.

When a study is conducted to determine whether an occupational exposure is associated with a disease, with no specific interest in the dose-response relationship, and when the estimate of the association is large, adjusted and unadjusted estimates are often similar. This has been shown in studies of occupational lung cancer risk, for which smoking is a strong potential confounder. In particular, using an indirect approach that is a type of sensitivity analysis, Axelson calculated that the confounding effect of smoking can hardly explain relative risks greater than 1.5 or below 0.7 when national rates are used for comparison (Axelson 1978; Axelson and Steenland 1988). He made sound assumptions about the proportions of moderate smokers (40%) and heavy smokers (10%) in the population used for calculating the number of expected cases and also about the effects of moderate smoking (relative risk, 10) and of heavy smoking (relative risk, 20) on lung cancer risk. Then, the adjusted relative risks were calculated for different scenarios of association between smoking and being employed in specific occupations (Table 42.2). Axelson's suggestion that, under common circumstances, strong risk factors have weak confounding effects was investigated further and supported (Gail et al. 1988; Siemiatycki et al. 1988; Flanders and Khoury 1990).

In developing a protocol for a case-control study on the risk of female breast cancer associated with occupational exposure to magnetic fields, a simulation study was carried out to evaluate the potential confounding effects of several risk factors (Goodman et al. 2002). Twelve potential confounders, including a family history of breast cancer, country of birth, age at menopause, and obesity, were selected on the basis of recent reviews on breast cancer epidemiology and evaluated both in

Table 42.2 Risk ratios for lung cancer in relation to the fraction of smokers in various hypothetical populations (Source: Axelson 1978; Axelson and Steenland 1988)

	Type of smokers in the population (percentages)		
Non-smokers (risk of 1)	Moderate smokers (risk of 10)	Heavy smokers (risk of 20)	Rate ratio
100	–	–	0.15
80	20	–	0.43
70	30	–	0.57
60	35	5	0.78
50[a]	40[a]	10[a]	1.00[a]
40	45	15	1.22
30	50	20	1.43
20	55	25	1.65
10	60	30	1.86
–	–	100	3.08

[a] Compared to reference population with 50% non-smokers, 40% moderate smokers, and 10% heavy smokers

univariable analyses and with combinations of two to five risk factors. Estimates of the strength of the associations between the risk factors and breast cancer risk and the prevalences of the risk factors in the general population were obtained from the literature. The aim was to identify confounders that, under different scenarios of their prevalence among cases, could increase a true odds ratio of 1 up to a distorted value of 1.5. In the univariable analysis, no risk factor was a strong confounder, unless an unrealistic increase in its prevalence among occupationally exposed women was assumed. Interestingly, the scenario in which the prevalences of several risk factors were increased also did not have a strong confounding effect. For instance, a twofold increase among exposed women in the prevalence of first-degree relatives with breast cancer, a history of cancer in one breast, benign proliferative breast disease, obesity, and consumption of at least two drinks per day inflated the odds ratio from unity to 1.38.

The similarity between adjusted and unadjusted estimates has also been shown empirically, in both cohort and case-control studies. *SMR*s of cancers of the lung, bladder, and intestine, unadjusted for smoking, strongly correlated with smoking-adjusted estimates in analyses of occupational factors in a cohort of US veterans (Blair et al. 1985). Analogously, smoking was found to be a weak confounder in a review of several occupational case-control studies on lung cancer (Simonato et al. 1988). When selecting the final model for analyzing a case-control study on occupational factors and lung cancer risk in two areas of Italy in 1990–1992 (Richiardi et al. 2004), we evaluated several models for addressing smoking as a confounder. Table 42.3 shows the results of an evaluation for two occupational categories, one positively associated and the other negatively associated with smoking. The evaluation showed that a simple model in which smokers are classified as current, former, and never can accommodate for most of the potential confounding effect.

Table 42.3 Odds ratio and 95% confidence intervals of lung cancer for two selected job titles, obtained using seven different methods to model smoking in an Italian case-control study on lung cancer (Source of data: Richiardi et al. 2004)[a]

	Retail trade salesmen (54 exposed cases)	Mail distribution clerks (58 exposed cases)
Model	*OR* (95% CI)[b]	*OR* (95% CI)
1	1.56 (1.04–2.35)	1.47 (1.00–2.17)
2	1.41 (0.93–2.15)	1.62 (1.07–2.45)
3	1.30 (0.85–2.01)	1.65 (1.08–2.52)
4	1.30 (0.84–2.03)	1.63 (1.06–2.51)
5	1.26 (0.81–1.95)	1.70 (1.10–2.61)
6	1.28 (0.82–2.00)	1.70 (1.10–2.63)
7	1.26 (0.81–1.98)	1.70 (1.10–2.65)

[a] All models were adjusted for age and study area. Model 1: no smoking variables; Model 2: smoking status categorized as ever/never smoker; Model 3: smoking status categorized as current/former (since at least 2 years)/never smoker; Model 4: same as model 3 with three levels for current smokers: 1–9, 10–19, 20+ pack-years; Model 5: same as model 3 with number of pack-years introduced as a continuous variable; Model 6: same as model 5 with 4 levels for time since cessation: 2–5, 6–10, 11–15, 16+ years; Model 7: same as model 6, using b-spline cubic regression with knots at 10, 20, 30, and 40 pack-years to model the cumulative number of cigarettes smoked

[b] *OR*, odds ratio adjusted for age, study area; CI, confidence interval

42.4.2 Healthy Worker Effect

Workers are not a random sample of the general population as the employment status is positively associated with the health status. First, relatively healthier people are more likely to seek a job and to be hired. Second, as sick people tend to leave their jobs, healthier workers remain employed longer. The two health-related selection forces cause a well-known selection bias in occupational epidemiology, known as the "healthy worker effect" (Fox and Collier 1976; McMichael 1976). The first phenomenon is known as the "healthy hire effect," whereas the second, associated with duration of employment, is known as the "healthy worker survivor effect" (Arrighi and Hertz-Picciotto 1994). The magnitude of the phenomena depends on the type of work, general social conditions (e.g., unemployment rate), the disease under study (e.g., studies of cancer are generally less biased than studies of diseases with shorter induction period), and the study design (Choi 1992). The healthy worker effect is also seen as a traditional confounding problem, as employment status is associated at the same time with the health status of workers and disease risk (Checkoway et al. 2004). Because of the healthy worker effect, cohorts of workers may have lower mortality rates than the general population. Negative results in occupational epidemiological studies may therefore hide harmful exposures. Moreover, an increase in risk of a disease may artificially plateau at the highest level of cumulative exposure, at which workers have the longest duration of employment (Stayner et al. 2003).

A logical approach for controlling, or at least decreasing, the bias introduced by the healthy worker effect is to use an appropriate internal or external comparison group, namely, a group of unexposed workers who possibly underwent similar health-related selection at the time of employment. Use of such a comparison group does not, however, imply unbiased estimates, as the healthy worker survivor effect may still persist. Indeed, as reviewed by Checkoway and colleagues (2004), four time-related factors should be considered: age at first employment, duration of employment, length of follow-up (members of a cohort can be followed-up also after they have left the job), and age at risk. Arrighi and Hertz-Picciotto (1996) evaluated four suggested methods for controlling the healthy worker effect: (1) restricting analyses to long-term survivors; (2) excluding recent exposures, introducing a lag of 10–20 years; (3) introducing current employment status as a confounder in the models; and (4) modeling employment status simultaneously as a confounder (the same as in the third approach) and as an intermediate time-dependent variable (if the risk factor for the disease under study is also a determinant of job termination and, therefore, of change in employment status). The latter technique uses the so-called G-method as suggested by Robins and colleagues (1986, 1992). This approach has the strongest theoretical support and was considered the most appropriate after empirical evaluation, although there are difficulties in its implementation. Lagging exposure is a valid, straightforward alternative that can be implemented when the induction period between exposure and disease is not short.

Case-control and cross-sectional studies are not free from the healthy worker effect. In case-control studies, it can result in differential sampling of controls from the exposed and the unexposed population. For instance, if controls are selected from hospitalized patients and individuals with a particular occupational exposure tend to be healthier, then the proportion of exposed controls is artificially decreased. The odds ratio would, therefore, be overestimated, a bias that reverses the usual underestimation of SMRs introduced by the healthy worker effect in cohort studies.

In a cross-sectional study, workers with higher exposure may have a paradoxically lower prevalence of diseases or symptoms known to be associated with exposure, because diseased workers would tend to leave jobs entailing the exposure.

42.4.3 Dose-Response Analysis

As discussed in chapter ▶Exposure Assessment of this handbook), the dose is the level of the risk factor at the target organ, while exposure refers to the level of the risk factor in the external environment. Although the dose is the biologically relevant measure, the amount of exposure, as a surrogate of the dose, is usually the only available information in occupational studies, so that a dose-response analysis is in fact an exposure-response analysis. In some studies, the actual dose can be estimated from measurements of exposure and knowledge of the specific agent uptake and clearance (US Environmental Protection Agency 2002).

Exposure can be measured using different metrics, namely duration, intensity, and cumulative level (cf. chapter ▶Exposure Assessment of this handbook). The selection of the metric should be based on the – often unknown – mechanism of disease development and on the nature of the exposure itself. Importantly, the choice of the metric influences the magnitude of the estimates and the shape of the dose-response (Blair and Stewart 1992). Cumulative exposure, that is, the product of intensity and duration, is a correct metric for several types of diseases where risk is directly proportional to dose. Duration of employment is a valid surrogate for cumulative exposure when intensity of exposure has been relatively constant over time, through working areas of the plant and across tenures (Checkoway 1986). Peak exposure is more important than duration in the study of diseases for which a threshold exists, such as back pain or acute toxicity.

A dose-response analysis is commonly carried out in occupational epidemiology for at least three main reasons. First, occupational exposures are time- and place-specific, implying that assessment of an association between an occupational exposure and a disease necessarily takes the level of exposure into account. Second, a dose-response relation is one of the well-known Bradford Hill's criteria for establishing causality (Hill 1965). When the risk for a disease increases continuously with increasing exposure, whatever the shape of the trend, the likelihood of a causal association is higher. However, on the one hand, a dose-response relation does not prove causality; on the other hand, the lack of such a relation does not imply lack of a causal association, as clearly demonstrated by threshold phenomena. Third, the dose-response analysis is one of the steps in risk assessment, which aims at quantifying the health effects of environmental and occupational exposures that can be modified by new policies and technologies. Risk assessment comprises (1) hazard identification, on the basis of evaluation of the available evidence on the health effects of the agent; (2) exposure assessment, identifying the nature of the exposure in the population, the characteristics of the exposed individuals, and the behavior of the agent in humans; (3) identification of the exposure-risk model, which implies a dose-response assessment; and (4) risk characterization, determining the exposure level-specific health effects in the population (Nurminen et al. 1999; Checkoway et al. 2004).

Often, data on exposure in occupational epidemiology are summarized as qualitative or semiquantitative indices. For instance, JEMs usually produce indices of intensity and probability of exposure on an ordinal scale. Such information offers little basis for a dose-response analysis. In other instances, if quantitative information is available, cumulative exposure can be estimated for each study subject; it must be born in mind, however, that quantitative estimates are affected by measurement errors, falling in the two broad categories of classical and Berkson error already discussed in Sect. 42.2.3. Among many possible examples, we cite here the dose-response analysis carried out by Steenland and colleagues (1998) on occupational exposure to diesel exhaust in the trucking industry and the risk for lung cancer, which was used for quantitative risk assessment by the Health Effects Institute (1999). Steenland and colleagues conducted a case-control study, obtaining information on lifetime work histories from interviews with the study subjects' next-of-kin and from retirement registries. Then, for the purpose of exposure assessment, workers were assigned to the category in which they had been employed the longest. Contemporaneously, an industrial hygiene survey was conducted to measure levels of exposure to elemental carbon (a marker of exposure to diesel exhaust, which is a complex mixture of gases and particulates) in the main job categories within the trucking industry (Zaebst et al. 1991). Combining the lifetime work histories with the results of the survey and making several assumptions, in particular with regard to past exposure, the cumulative exposure of each worker was estimated. Although quantitative data were obtained, the level of misclassification of exposure was still presumably high, albeit non-differential. In particular, each subject's true exposure in each job category was a random variation of the exposure level that was assigned to that job based on the industrial hygiene survey.

There are several approaches to dose-response analysis, including simple parametric models, categorical analysis, biological-based models, polynomial regression, spline regression, and non-parametric models, such as generalized additive models (cf. chapters ▶Analysis of Continuous Covariates and Dose-Effect Analysis and ▶Regression Methods for Epidemiological Analysis of this handbook). We will not describe and compare these techniques, but we highlight some aspects that are specific to occupational epidemiology. Interested readers may refer to the above chapters or one of the several available thematic textbooks (Härdle 1990; Hastie and Tibshirani 1990).

Categorical analysis, in which the exposure variable is subdivided into a certain number of categories on the basis of cut-points chosen a priori, is usually the starting point for a dose-response assessment, as it allows researchers to observe the shape of the dose-response relationship. The shape is obviously strongly influenced by the choice and number of the cut-points that can be decided upon according to biological considerations, if available, or other criteria, including established standards or the percentile method. Evidence that an association is limited to the highest exposure levels should not lead to disregard causality without careful consideration of the possibility of a threshold for the effect of interest.

When the exposure variable is continuous, the simplest approach consists in fitting a regression model with a term for the exposure (e.g., cumulative exposure). This implies assuming a priori a shape of the dose-response curve that seldom reflects biological knowledge, if any is available. When the exposure variable is not transformed, the assumed shape is usually log-linear or logistic. In occupational epidemiology, a leveling off in the increasing trend in risk for chronic diseases is often observed at the highest levels of exposure (Stayner et al. 2003). The explanations for this leveling off can be either biological (e.g., saturation phenomena, depletion of susceptible individuals) or methodological (e.g., misclassification of exposure, healthy worker effect), and a log transformation of the cumulative exposure variable is an option to consider.

Among the more complex alternatives, spline regression and its variants (b-splines and loess among the most popular) can be implemented quite easily with common software packages. It is therefore being used increasingly in occupational and environmental epidemiology (Greenland et al. 2000; Steenland et al. 2001; Thurston et al. 2002; Steenland and Deddens 2004). Spline regression, which is based on piecewise polynomials, has the advantage of providing a smoothed dose–response curve, although it does not always produce easily interpretable estimates (Harrell et al. 1988; Greenland 1995). Figure 42.2 shows an example of a dose-response analysis of data from men included in a case-control study on occupational factors and lung cancer risk that we carried out in two areas of northern Italy in

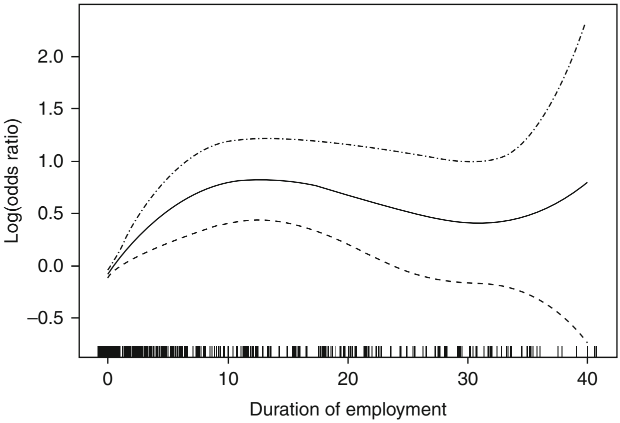


Fig. 42.2 Association between duration of employment in occupations known to entail exposure to lung carcinogens and risk of lung cancer modeled using a generalized additive model with cubic b-splines (four degrees of freedom), adjusted for study area, age, and cigarette smoking (current/former/never smokers) (Source of data: Richiardi et al. 2004)

1990–1992 (Richiardi et al. 2004). The odds ratio (plotted on the log scale) of lung cancer increased with duration of employment until 10–15 years and slightly decreased after that. Estimates for the durations of employment above 30 years are not interpretable because of few observations and consequent large confidence intervals. This shape in dose-response is not entirely unexpected when duration of exposure is used in the analysis, as subjects with longer duration of employment may be those with lower intensity of exposure and better health status.

42.5 Primary Prevention

How can occupational epidemiology help evaluate the need and effectiveness of primary prevention interventions and policies?

Sound epidemiological studies are typically needed to produce evidence of toxicity for occupational agents when long-term effects are present such as in occupational cancer that we use here as an example (Merletti and Mirabelli 2004). Complex mixtures entailing occupational exposures were among the first causes of cancer to be identified and finally led to the identification of specific causal agents. Thus, the study of occupational cancers offered precious insights and paradigms for occupational epidemiology at large.

Agents currently established as causes of occupational cancer and occupations with sufficient evidence of increased cancer risk according to the International Agency for Research on Cancer can be found in this textbook (see chapter ▶Cancer Epidemiology of this handbook; IARC 2011).

In appraising the body of evidence on occupational hazards and its relevance for control of occupational exposures, consideration must be given to the problem of who should bear the burden of proof and what the proof should consist of: whether evidence of benefit from intervention or evidence of harm from exposure. Occupational exposures are imposed upon individuals who have little, if any, personal choice, freedom, and responsibility in accepting or avoiding them. Furthermore, they often lack the basic necessary knowledge. As consequence, the burden of proof is on the employer, to demonstrate that the production process is safe. Evidence that exposure may be harmful is sufficient to require intervention to eliminate it.

Primary prevention, in the field of exposure to carcinogens as well as of other chemical and physical hazards at work, is based on the application of basic industrial hygiene strategies at the industry level: (1) substitution with agents intended not to be as dangerous, (2) fully enclosed processing, and (3) strict control of exposure by reduction of amounts used, by local exhaust, by personal protection, by cleaning practices, etc. This means to reduce the number of potentially exposed workers and their exposure level. Exposure control is better implemented by embedding it in the project of plants and processes, aiming to workers' protection as well as to that of neighboring communities.

At the community and country level, primary prevention entails adopting regulations intended to favor preventive measures or to enforce them. The first country to forbid the manufacture of certain chemicals because of their carcinogenicity was the United Kingdom, with the Carcinogenic Substances Regulations in 1967, prohibiting beta-naphthylamine, benzidine, 4-aminobiphenyl, and 4-nitrobiphenyl (UK 1967). The EC regulation on carcinogens at work has been developed starting with the 90|394|EEC Directive, but still today the only carcinogenic agents whose production and use is forbidden, apart from asbestos, are the same four as in the UK Carcinogenic Substances Regulations. In the USA, no formal ban has been put on any carcinogenic agent, production, or process on grounds of workers' protection. Permissible exposure levels (PELs) have been established by the Occupational Safety and Health Administration (OSHA) largely on the basis of the 1987 list of the American Conference of Governmental Industrial Hygienists (ACGIH) threshold limit values (TLVs), with the result that (1) the TLV's list has been updated and expanded by ACGIH, but the list of PELs is unchanged, and (2) certain agents are commonly recognized carcinogens, but their PELs were established without taking their cancer-causing properties into account (Smith and Mendeloff 1999).

Despite these limitations, OSHA and EPA in the USA and the EC in its regulation on classification, labeling and packaging of dangerous substances publish lists of substances officially recognized as carcinogens. The availability of lists of carcinogenic, and in general of toxic, chemicals is a useful tool for hazard identification, even if limited to intentionally used agents.

Workers' information on their exposures and on the risks entailed by them is a fundamental issue. It is the first step in their empowerment to verify that appropriate measures have been taken. The EC regulation requires that specific information is given to exposed workers, including special instructions on how to deal with accidents and emergencies.

Provided local regulations have been adopted, like all EC Member States should have done, law enforcement through technical public services specialized in inspecting workplaces is another key issue. Further, workers should be able to stand in courts not only when they are affected by work-related conditions but just because they are exposed, and their cases should be fairly settled, which does not seem to occur currently even in large EC Member States (Editorial 2003).

It may be surprising that systematic reviews on the effectiveness of interventions and/or implementation activities aimed at exposure control are generally lacking. In the area of occupational cancer, an exception may be the review by Kogevinas and coworkers in 1998 on the rubber industry (Kogevinas et al. 1998), where some changes in overall technology and chemistry were considered along with evidence on the persistence of previously observed cancer risks. This review is useful to point out the many and different difficulties we are confronted with while trying to gather evidence of effectiveness in occupational cancer prevention:

1. The long induction period of most human cancers prevents driving conclusions from early observations after changes are introduced, since workers first employed after intervention are not yet at risk, or fully at risk, of developing the disease.
2. Longer-term observations, however, are difficult to carry out; they are also difficult to interpret because of changing patterns of incidence/mortality in the disease of interest, and of possible complex interactions with other exposures.
3. Often, the exposure characteristics are not well understood and recorded, so it may become impossible to assess the quantitative relationship between exposure level and disease occurrence, which is precisely what is needed when exposure levels are reduced, but the agent is not completely eliminated.
   (a) Sometimes, the nature of the relevant exposure is not understood, so that a carcinogenic agent may be withdrawn, but its substitutes may be as dangerous, or almost as dangerous.
   (b) Both industry-based and community-based epidemiological studies have major limits in exposure assessment, due to lack of suitable exposure data, and this is the origin of major uncertainties and controversies in the interpretation of epidemiological evidence.

This picture explains why it is difficult to obtain evidence of cancer risk reduction following the adoption of control measures and why reports of this kind of evidence are rare.

Within the limits of the above-mentioned uncertainties, some widespread occupational cancer risks (Cruickshank and Squire 1950) seem to have disappeared from industrial and agricultural settings in Europe and in the USA. Furthermore, some carcinogenic exposures also disappeared or have been reduced to lower levels in developed countries. The subsequent reduction of the fraction of cancers attributable to occupation can be estimated, provided adequate data are available (Armstrong and Darnton 2008), and has to be the object of future scientific investigations.

Some contradictory experiences occurred either: Agents have been substituted with others now seemingly entailing the same risks, carcinogenic contaminants have been eliminated from agents used in certain industries only to be introduced in agents used in other processes, and only partial elimination of risk has been achieved when relevant exposures were to complex mixtures rather than to simple chemicals (Evanoff et al. 1993). Therefore, workers' exposure to carcinogens in industrialized countries is still not controlled as completely as it should be, given our current knowledge of the carcinogenic properties of chemical and physical agents. The most critical point, however, is continuation of productions and processes entailing exposure to carcinogens in developing countries, often lacking experience in the management of industrial hazards and power to enforce sound control strategies (Jeyaratnam 1994).

42.6 Conclusions

Attempts have been made to estimate the global burden of disease and injury due to occupational factors (Leigh et al. 1997, 1999; Ezzati et al. 2002; Rushton et al. 2010). Although such global statistics are of difficult interpretation, given the very large number of assumptions underlying them, two major conclusions can be drawn: (1) The problem is still an important one throughout the world, including developed regions; (2) the burden is shifting to the developing world, which accounts to 70% of the world's workforce and where the globalization of industry is resulting in increased exposure to occupational agents. The situation is exacerbated by unsafe technology, transfer of hazardous industries and wastes from developed to developing countries, use of agents banned or restricted elsewhere, poor health and nutritional status of the workforce, and ineffective legislation on occupational safety and health. Although prevention of exposure to occupational hazards will come from political and economic changes in the world, just as political and economic interests are the determinants of the present situation, much can still achieved, even in the current international situation (Pearce et al. 1994).

The applications of occupational epidemiology in public health decision-making are broadening, providing inputs to risk assessment, evaluation of occupational guidelines, and extrapolation of findings from occupational settings to communities with the aim of setting policies at population level. These multiple applications mean increasing responsibility to ensure ethical scientific conduct and clear, thorough communication of the assumptions, limitations, and uncertainties, of the results of research and of risk assessment (Kriebel and Tickner 2001).

Recent discoveries in molecular biology and genetics have made it possible for researchers to examine how genetic characteristics affect responses to occupational and environmental exposures. The use of genetic biomarkers in epidemiology has provided potential understanding of the underlying mechanisms of disease and therefore ultimately contributes to public health. Despite the potential benefits of genetic information, its collection in epidemiological studies, particularly in occupational settings, presents ethical, legal, and social challenges. Clarifying gene-environment interactions will have implications for difficult regulatory questions, such as protecting the most susceptible members of the population and its subgroups, but in the case of workers, genetic information could be used to discriminate them (Christiani et al. 2001). The challenge of identifying and applying genetic information in the study of human diseases in instances in which it will make a difference to prevention and public health (Millikan 2002; Merikangas and Risch 2003; Schulte 2004) may well also apply to occupational epidemiology.

Acknowledgements This work was partially supported by the Italian Association for Cancer Research and the Special Project Oncology, Compagnia San Paolo. We thank Francesco Barone Adesi for useful comments.

References

Ahrens W (1999) Retrospective assessment of occupational exposure in case control studies. Ecomed, Landsberg

Ahrens W, Jöckel KH, Brochard P, Bolm-Audorff U, Grossgarten K, Iwatsubo Y, Orlowski E, Pohlabeln H, Berrino F (1993) Retrospective assessment of asbestos exposure–I. Case-control analysis in a study of lung cancer: efficiency of job-specific questionnaires and job exposure matrices. Int J Epidemiol S2:83–95

Andersen A, Barlow L, Engeland A, Kjaerheim K, Lynge E, Pukkala E (1999) Work related cancer in the Nordic countries. Scand J Work Environ Health 25:1–116

Armstrong BG (1998) Effects of measurement error on epidemiological studies of environmental and occupational exposures. Occup Environ Med 55:651–656

Armstrong BG, Darnton A (2008) Estimating reduction in occupational disease burden following reduction in exposure. Occup Environ Med 65:592–596

Arrighi HM, Hertz-Picciotto I (1994) The evolving concept of the healthy worker effect. Epidemiology 5:189–196

Arrighi HM, Hertz-Picciotto I (1996) Controlling the healthy worker survivor effect: an example of arsenic exposure and respiratory cancer. Occup Environ Med 53:455–462

Axelson O (1978) Aspects of confounding in occupational health epidemiology. Scand J Work Environ Health 4:85–89

Axelson O, Steenland K (1988) Indirect methods of assessing the effects of tobacco use in occupational studies. Am J Ind Med 13:105–118

Berrino F, Richiardi L, Boffetta P, Esteve J, Belletti I, Raymond L, Troschel L, Pisani P, Zubiri L, Ascunce N, Guberan E, Tuyns A, Terracini B, Merletti F, Milan JEM Working Group (2003) Occupation and larynx and hypopharynx cancer: a job-exposure matrix approach in an international case-control study in France, Italy, Spain and Switzerland. Cancer Causes Control 14:213–223

Blair A, Stewart PA (1992) Do quantitative exposure assessments improve risk estimates in occupational studies of cancer? Am J Ind Med 21:53–63

Blair A, Hoar SK, Walrath J (1985) Comparison of crude and smoking-adjusted standardized mortality ratios. J Occup Med 27:881–884

Boffetta P, Saracci R, Andersen A, Bertazzi PA, Chang-Claude J, Cherrie J, Ferro G, Frentzel-Beyme R, Hansen J, Olsen J, Plato N, Teppo L, Westernholm P, Winter PD, Zocchetti C (1997) Cancer mortality among man-made vitreous fiber production workers. Epidemiology 8: 259–268

Boffetta P, Andersen A, Hansen J, Olsen J, Plato N, Teppo L, Westernholm P, Saracci R (1999) Cancer incidence among European man-made vitreous fiber production workers. Scand J Environ Health 25:222–226

Boffetta P, Richiardi L, Berrino F, Esteve J, Pisani P, Crosignani P, Raymond L, Zubiri L, Del Moral A, Lehmann W, Donato F, Terracini B, Tuyns A, Merletti F (2003) Occupation and larynx and hypopharynx cancer: an international case control study in France, Italy, Spain, and Switzerland. Cancer Causes Control 14:203–212

Bouyer J, Dardenne J, Hemon D (1995) Performance of odds ratios obtained with a job-exposure matrix and individual exposure assessment with special reference to misclassification errors. Scand J Work Environ Health 21:265–271

Boyd JT, Doll R, Faulds JS, Leiper J (1970) Cancer of the lung in iron (haematite) miners. Br J Ind Med 27:97–105

Brüske-Hohlfeld I, Möhner M, Pohlabeln H, Ahrens W, Bolm-Audorff U, Kreienbrock L, Kreuzer M, Jahn I, Wichmann HE, Jöckel KH (2000) Occupational lung cancer risk for men in Germany: results from a pooled case-control study. Am J Epidemiol 151:384–395

Burstyn I, Kromhout H, Kauppinen T, Heikkala P, Boffetta P (2000) Statistical modelling of the determinants of historical exposure to bitumen and polycyclic aromatic hydrocarbons among paving workers. Ann Occup Hyg 44:43–56

Burstyn I, Boffetta P, Kauppinen T, Heikkala P, Svane O, Partanen T, Stucker I, Frentzel-Beyme R, Ahrens W, Merzenich H, Heederik D, Hooiveld M, Langard S, Randem BG, Jarvholm B, Bergdahl I, Shaham J, Ribak J, Kromhout H (2003) Estimating exposures in the asphalt industry for an international epidemiological cohort study of cancer risk. Am J Ind Med 43:3–17

Carter T (2000) Diseases of occupations – a short history of their recognition and prevention. In: Baxter PJ, Adams PH, Tar-Ching Aw, Cockcroft A, Harrington JM (eds) Hunter's diseases of occupations. Arnold, London, pp 917–925

Case RAM, Hosker ME (1954) Tumours of the urinary bladder as an occupational disease in the rubber industry in England and Wales. Br J Prev Soc Med 8:39–50

Case RAM, Hosker ME, McDonald DB (1954) Tumours of the urinary bladder in workmen engaged in the manufacture and use of certain dyestuff intermediates in the British chemical industry. Br J Ind Med 11:75–104

Checkoway H (1986) Methods of treatment of exposure data in occupational epidemiology. Med Lav 77:48–73

Checkoway H, Pearce N, Kriebel D (2004) Research methods in occupational epidemiology. Oxford University Press, Oxford

Choi BCK (1992) Definition, sources, magnitude, effect modifiers, and strategies of reduction of the healthy worker effect. J Occup Med 34:979–988

Christiani DC, Sharp RR, Collman GW, Suk WA (2001) Applying genomic technologies in environmental health research: challenges and opportunities. J Occup Environ Med 43:526–533

Coggon D (2000) Estimating the extent of occupational injuries and disease. In: Baxter PJ, Adams PH, Tar-Ching Aw, Cockcroft A, Harrington JM (eds) Hunter's diseases of occupations. Arnold, London, pp 27–35

Corbin M, Maule M, Richiardi L, Simonato L, Merletti F, Pearce N (2008) Semi-Bayes and empirical Bayes adjustment methods for multiple comparisons. Epidemiol Prev 32:108–110

Costa G, Merletti F, Segnan N (1989) A mortality cohort study in a north Italian aircraft factory. Br J Ind Med 46:738–743

Cruickshank CND, Squire JR (1950) Skin cancer in the engineering industry from the use of mineral oil. Brit J Ind Med 7:1–11

de Vocht F, Vermeulen R, Burstyn I, Sobala W, Dost A, Taeger D, Bergendorf U, Straif K, Swuste P, Kromhout H; EU-EXASRUB consortium (2008) Exposure to inhalable dust and its cyclohexane soluble fraction since the 1970s in the rubber manufacturing industry in the European Union. Occup Environ Med 65:384–391

Doll R (1952) The causes of death among gas-workers with special reference to cancer of the lung. Br J Ind Med 9:180–185

Doll R (1955) Mortality from lung cancer in asbestos workers. Br J Ind Med 12:81–86

Doll R (1975) Pott and the prospects for prevention. Br J Cancer 32:263–74

Dosemeci M, Stewart PA, Blair A (1990a) Three proposals for retrospective, semi-quantitative exposure assessments and their comparison with other assessment methods. Appl Occup Environ Hyg 5:52–59

Dosemeci M, Wacholder S, Lubin JH (1990b) Does non differential misclassification of exposure always bias a true effect toward the null value? Am J Epidemiol 132:746–748

Dosemeci M, Li GL, Hayes RB, Yin SN, Linet M, Chow WH, Wang YZ, Jiang ZL, Dai TR, Zhang WU, Chao XJ, Ye PZ, Kou QR, Fan YH, Zhang XC, Lin XF, Meng JF, Zho JS, Wacholder S, Kneller R, Blot WJ (1994a) Cohort study among workers exposed to benzene in China: II. Exposure assessment. Am J Ind Med 26:401–411

Dosemeci M, McLaughlin JK, Chen JQ, Hearl F, McCrawley M, Wu Z, Chen RG, Peng KL, Chen AL, Rexing SH (1994b) Indirect validation of a retrospective method of exposure assessment used in a nested case-control study of lung cancer and silica exposure. Occup Environ Med 51:136–138

Dosemeci M, Yin SN, Linet M, Wacholder S, Rothman N, Li GL, Chow WH, Wang YZ, Jiang ZL, Dai TR, Zhang WU, Chao XJ, Ye PZ, Kou QR, Fan YH, Zhang XC, Lin XF, Meng JF, Zho JS, Blot WJ, Hayes RB (1996) Indirect validation of benzene exposure assessment by association with benzene poisoning. Environ Health Perspect S6:1343–1347

Editorial (2003) Who will take responsibility for corporate killing? Lancet 361:1921

Evanoff BA, Gustavsson P, Hogstedt C (1993) Mortality and incidence of cancer in a cohort of Swedish chimney sweeps: an extended follow-up study. Br J Ind Med 50:450–459

Ezzati M, Lopez AD, Rodgers A, Vander Hoorn S, Murray CJ, Comparative Risk Assessment Collaborating Group (2002). Selected major risk factors and global and regional burden of disease. Lancet 360:1347–1360

Flanders WD, Khoury MJ (1990) Indirect assessment of confounding: graphic description and limits on effect of adjusting for covariates. Epidemiology 1:239–246

Fox AJ, Collier PF (1976) Low mortality rates in industrial cohort studies due to selection for work and survival in the industry. Br J Prev Soc Med 30:225–230

Fritschi L, Nadon L, Benke G, Lakhani R, Latreille B, Parent ME, Siemiatycki J (2003) Validation of expert assessment of occupational exposures. Am J Ind Med 43:519–522

Gail MH, Wacholder S, Lubin JH (1988) Indirect corrections for confounding under multiplicative and additive risk models. Am J Ind Med 13:119–130

Goodman M, Kelsh M, Ebi K, Iannuzzi J, Langholz B (2002) Evaluation of potential confounders in planning a study of occupational magnetic field exposure and female breast cancer. Epidemiology 13:50–58

Grayson JK (1996) Radiation exposure, socioeconomic status, and brain tumor risk in the US Air Force: a nested case-control study. Am J Epidemiol 143:480–486

Greaves IA, Eisen EA, Smith TJ, Pothier LJ, Kriebel D, Woskie SR, Kennedy SM, Shalat S, Monson RR (1997) Respiratory health of automobile workers exposed to metal-working fluid aerosols: respiratory symptoms. Am J Ind Med 32:450–459

Greenland S (1995) Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology 6:356–365

Greenland S, Poole C (1994) Empirical-Bayes and semi-Bayes approaches to occupational and environmental hazard surveillance. Arch Environ Health 49:9–16

Greenland S, Sheppard AR, Kaune WT, Poole C, Kelsh MA (2000) A pooled analysis of magnetic fields, wire codes, and childhood leukemia. Childhood Leukemia- EMF Study Group. Epidemiology 11:624–634

Haldorsen T, Andersen A, Boffetta P (2004) Smoking-adjusted incidence of lung cancer by occupation among Norwegian men. Cancer Causes Control 15:139–147

Harber P, Muranko H, Shvartsblat S, Solis S, Torossian A, Oren T (2003) A triangulation approach to historical exposure assessment for the carbon black industry. J Occup Environ Med 45: 131–143

Härdle W (1990) Applied nonparametric regression. Cambridge University Press, Cambridge/ New York

Harrell FE Jr, Lee KL, Pollock BG (1988) Regression models in clinical studies: determining relationships between predictors and response. J Natl Cancer Inst 80:1198–1202

Hastie T, Tibshirani R (1990) Generalized additive models. Chapman and Hall, London

Hayes RB, Yin SN, Dosemeci M, Li GL, Wacholder S, Travis LB, Li CY, Rothman N, Hoover RN, Linet MS (1997) Benzene and the dose-related incidence of Hematologic Neoplasms in China. J Natl Cancer Inst 89:1065–1071

Health Effects Institute (1999) Diesel emissions and lung cancer: epidemiology and quantitative risk assessment. A Special Report of the Institute's Diesel Epidemiology Expert Panel. Flagship Press, Andover

Hill AB (1965) The environment and disease: association or causation. Proc R Soc Med 58: 295–300

Hoar SK, Morrison AS, Cole P, Silverman DT (1980) An occupational exposure linking system for the study of occupational carcinogenesis. J Occup Med 22:722–726

IARC (2011) List of classifications by alphabetical order. http://monographs.iarc.fr/ENG/Classification/index.php. Accessed 6 Dec 2011

Iwatsubo Y, Pairon JC, Boutin C, Menard O, Massin N, Caillaud D, Orlowski E, Galateau Salle F, Bignon J, Brochard P (1998) Pleural mesothelioma: dose response relation at low levels of asbestos exposure in a French population based case-control study. Am J Epidemiol 148: 133–142

Jeyaratnam J (1994). Transfer of hazardous industries. In: Pearce N, Matos E, Vainio H, Boffetta P, Kogevinas M (eds) Occupational cancer in developing countries. International Agency for Research on Cancer, Lyon, pp 23–29

Kauppinen T, Partanen T (1988) Use of plant- and period-specific job-exposure matrices in studies on occupational cancer. Scand J Work Environ Health 14:161–167

Kauppinen TP, Pannet B, Marlow DA, Kogevinas M (1994) Retrospective assessment of exposure through modelling in a study on cancer risks among workers exposed to phenoxy herbicides, chlorophenols and dioxins. Scand J Work Environ Health 20:262–271

Kauppinen T, Heikkilä P, Plato N, Woldbaek T, Lenvik K, Hansen J, Kristjansson V, Pukkala E (2009) Construction of job-exposure matrices for the Nordic Occupational Cancer Study (NOCCA). Acta Oncol 48:791–800

Kjaerheim K, Boffetta P, Hansen J, Cherrie J, Chang-Claude J, Eilber U, Ferro G, Guldner K, Olsen JH, Plato N, Proud L, Saracci R, Westerholm P, Andersen A (2002) Lung cancer among rock and slag wool production workers. Epidemiology 13:445–553

Kogevinas M, Sala M, Boffetta P, Kazerouni N, Kromhout H, Zahm SH (1998) Cancer risk in the rubber industry: a review of the recent epidemiological evidence. Occup Environ Med 55:1–12

Kriebel D, Tickner J (2001) Reenergizing public health through precaution. Am J Public Health 91:1351–1355

Kromhout H, Symanski E, Rappaport SM (1993) A comprehensive evaluation of within- and between-workers components of occupational exposure to chemical agents. Ann Occup Hyg 37:253–270

Larsen SB, Spano M, Giwercman A, Bonde JP (1999) Semen quality and sex hormones among organic and traditional Danish farmers. ASCLEPIOS Study Group. Occup Environ Med 56:139–144

Lavoué J, Gérin M, Coté J, Lapointe R (2007) Mortality and cancer experience of Quebec aluminium reduction plant workers. Part I: the reduction plants and coal tar pitch volatile (CTPV) exposure assessment. J Occup Environ Med 49:997–1008

Leigh JP, Marcowitz SB, Fahs M, Shin C, Landrigan PJ (1997) Occupational injury and illness in the United States. Arch Int Med 157:1557–1568

Leigh J, Macaskill P, Kuosma E, Mandryk J (1999) Global burden of disease and injury due to occupational factors. Epidemiology 10:626–631

Luce D, Gerin M, Berrino F, Pisani P, Leclerc A (1993) Sources of discrepancies between a job exposure matrix and a case by case expert assessment for occupational exposure to formaldehyde and wood-dust. Int J Epidemiol S2:113–120

Macaluso M, Vineis P, Continenza D, Ferrario F, Pisani P, Audisio R (1983) Job exposure matrices: experience in Italy. In: Acheson ED (ed) Job exposure matrices: proceedings of a conference held in April 1982 the University of Southampton. MRC-EEU Scientific Report no. 2, Southampton General Hospital

McMichael AJ (1976) Standardized mortality ratios and the "healthy worker effect": scratching beneath the surface. Occup Med 18:165–168

Merikangas KR, Risch N (2003) Genomic priorities and public health. Science 302:599–601

Merletti F, Mirabelli D (2004) Occupational exposures. In: Evidence-based cancer prevention: strategies for NGOs. A UICC handbook for Europe. International Union Against Cancer, Geneva

Merletti F, Boffetta P, Ferro G, Pisani P, Terracini B (1991) Occupation and cancer of the oral cavity/oropharynx in Turin, Italy. Scand J Work Environ Health 17:248–254

Miettinen OS, Wang JD (1981) An alternative to the proportionate mortality ratio. Am J Epidemiol 114:144–1488

Millikan R (2002) The changing face of epidemiology in the genomics era. Epidemiology 13:472–480

Nurminen M, Nurminen T, Corvalàn CF (1999) Methodologic issues in epidemiologic risk assessment. Epidemiology 10:585–593

Orlowski E, Pohlabeln H, Berrino F, Ahrens W, Bolm-Audorff U, Grossgarten K, Iwatsubo Y, Jöckel KH, Brochard P (1993) Retrospective assessment of asbestos exposure–II. At the job level: complementarity of job-specific questionnaire and job exposure matrices. Int J Epidemiol S2:96–105

Pairon JC, Orlowski E, Iwatsubo Y, Billon-Gailland MA, Dufour G, Chamming's S, Archambault C, Bignon J, Brochard P (1994) Pleural mesothelioma and exposure to asbestos: evaluation from work histories and analysis of asbestos bodies in bronchoalveolar lavage fluid or lung tissue in 131 patients. Occup Environ Med 51:244–249

Pearce N, Matos E, Vainio H, Boffetta P, Kogevinas M (eds) (1994) Occupational cancer in developing countries. International Agency for Research on Cancer, Lyon, pp 23–29

Peters S, Talaska G, Jönsson BA, Kromhout H, Vermeulen R (2008) Polycyclic aromatic hydrocarbon exposure, urinary mutagenicity, and DNA adducts in rubber manufacturing workers. Cancer Epidemiol Biomarkers Prev 17:1452–1459

Peters S, Vermeulen R, Olsson A, Van Gelder R, Kendzia B, Vincent R, Savary B, Williams N, Woldbæk T, Lavoué J, Cavallo D, Cattaneo A, Mirabelli D, Plato N, Dahmann D, Fevotte J, Pesch B, Brüning T, Straif K, Kromhout H (2012) Development of an exposure measurement database on five lung carcinogens (ExpoSYN) for quantitative retrospective occupational exposure assessment. Ann Occup Hyg 56(1):70–79

Pohlabeln H, Wild P, Schill W, Ahrens W, Jahn I, Bolm-Audorff U, Jöckel KH (2002) Asbestos fibre years and lung cancer: a two phase case-control study with expert exposure assessment. Occup Environ Med 59:410–414

Potashnik G, Ben-Aderet N, Israeli R, Yanai-Inbar I, Sober I (1978) Suppressive effect of 1, 2-dibromo-3-chloropropane on human spermatogenesis. Fertil Steril 30:444–447

Preller L, van den Bosch LM, van den Brandt PA, Kauppinen T, Goldbohm A (2010) Occupational exposure to silica and lung cancer risk in the Netherlands. Occup Environ Med 67:657–663

Pukkala E, Aspholm R, Auvinen A, Eliasch H, Gundestrup M, Haldorsen T, Hammar N, Hrafnkelsson J, Kyyronen P, Linnersjo A, Rafnsson V, Storm H, Tveten U (2002) Incidence of cancer among Nordic airline pilots over five decades: occupational cohort study. BMJ 325: 567–569

Ramachandran G (2001) Retrospective exposure assessment using Bayesian methods. Ann Occup Hyg 45:651–667

Ramazzini B (1964) Diseases of workers (Translation of De Morbis Artificum 1713 text). Hafner, New York

Richardson DB, Ciampi A (2003) Effects of measurement error when an exposure variable is constrained by a lower limit. Am J Epidemiol 157:355–363

Richiardi L, Boffetta P, Simonato L, Forastiere F, Zambon P, Fortes C, Gaborieau V, Merletti F (2004) Occupational risk factors for lung cancer in men and women: a population-based case-control study in Italy. Cancer Causes Control 15:285–294

Robins JM (1986) A new approach to causal inference in mortality studies with sustained exposure period: application to control of the healthy worker survivor effect. Math Model 7:1393–1512

Robins JM, Blevins D, Ritter G, Wulfson M (1992) G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients. Epidemiology 3:319–336

Rödelsperger K, Jöckel KH, Pohlabeln H, Romer W, Woitowitz HJ (2001) Asbestos and man-made vitreous fibers as risk factors for diffuse malignant mesothelioma: results from a German hospital-based case-control study. Am J Ind Med 39:262–275

Rothman KJ, Greenland S, Lash TL (eds) (2008) Modern epidemiology, 3rd edn. LWW, Philadelphia

Rushton L, Bagga S, Bevan R, Brown TP, Cherrie JW, Holmes P, Fortunato L, Slack R, Van Tongeren M, Young C, Hutchings SJ (2010) Occupation and cancer in Britain. Br J Cancer 102:1428–1437

Rybicki BA, Cole Johnson C, Peterson EL, Kortsha GX, Gorell JM (1997) Comparability of different methods of retrospective exposure assessment of metals in manufacturing industries. Am J Ind Med 31:36–43

Sali D, Boffetta P, Andersen A, Cherrie JW, Claude JC, Hansen J, Olsen JH, Pesatori AC, Plato N, Teppo L, Westerholm P, Winter P, Saracci R (1999) Non neoplastic mortality of European workers who produce man made vitreous fibres. Occup Environ Med 56:612–617

Schulte PA (2004) Some implications of genetic biomarkers in occupational epidemiology and practice. Scand J Work Environ Health 30:71–79

Siemiatycki J, Day N, Fabry J, Cooper J (1981) Discovering carcinogens in the occupational environment: a novel epidemiologic approach. J Natl Cancer Inst 66:217–225

Siemiatycki J, Wacholder S, Dewar R, Cardis E, Greenwood C, Richardson L (1988) Degree of confounding bias related to smoking, ethnic group, and socioeconomic status in estimates of the associations between occupation and cancer. J Occup Med 30:617–625

Siemiatycki J, Fritschi L, Nadon L, Gerin M (1997) Reliability of an expert rating procedure for retrospective assessment of occupational exposures in community based case-control studies. Am J Ind Med 31:280–286

Simonato L, Vineis P, Fletcher AC (1988) Estimates of the proportion of lung cancer attributable to occupational exposure. Carcinogenesis 9:1159–1165

Smith JS, Mendeloff JM (1999) A quantitative analysis of factors affecting PELs and TLVs for carcinogens. Risk Anal 19:1223–1234

Stayner L, Steenland K, Dosemeci M, Hertz-Picciotto I (2003) Attenuation of exposure-response curves in occupational cohort studies at high exposure levels. Scand J Work Environ Health 29:317–324

Steenland K, Deddens JA (2004) Practical guide to dose-response analyses and risk assessment in occupational epidemiology. Epidemiology 15:63–70

Steenland K, Deddens JA, Stayner L (1998) Diesel exhaust and lung cancer in the trucking industry: exposure-response analyses and risk assessment. Am J Ind Med 34:220–228

Steenland K, Bray I, Greenland S, Boffetta P (2000a) Empirical Bayes adjustments for multiple results in hypothesis-generating or surveillance studies. Cancer Epidemiol Biomarkers Prev 9:895–903

Steenland K, Deddens JA, Zhao S (2000b) Biases in estimating the effect of cumulative exposure in log-linear models when estimated exposure levels are assigned. Scand J Work Environ Health 26:37–43

Steenland K, t'Mannetje A, Boffetta P, Stayner L, Attfield M, Chen J, Dosemeci M, De Klerk N, Hnizdo E, Koskela R, Checkoway H, International Agency for Research on Cancer (2001) Pooled exposure-response analyses and risk assessment for lung cancer in 10 cohorts of silica-exposed workers: an IARC multicentre study. Cancer Causes Control 12:773–784

Stengel B, Pisani P, Limasset JC, Bouyer J, Berrino F, Hémon D (1993) Retrospective evaluation of occupational exposure to organic solvents: questionnaire and job exposure matrix. Int J Epidemiol S2:72–82

Stewart P, Stewart W, Heineman EF, Dosemeci M, Linet M, Inskip PD (1996) A novel approach to data collection in a case-control study of cancer and occupational exposures. Int J Epidemiol 25:744–752

Stucker I, Boyer J, Mandereau L, Hémon D (1993) Retrospective evaluation of the exposure to polycyclic aromatic hydrocarbons: comparative assessments with a job exposure matrix and by experts in industrial hygiene. Int J Epidemiol S2:106–112 t'Mannetje A, Steenland K, Checkoway H, Koskela RS, Koponen M, Attfield M, Chen J, Hnizdo E, De Klerk N, Dosemeci M (2002) Development of quantitative exposure data for a pooled exposure-response analysis of 10 silica cohorts. Am J Ind Med 42:73–86

Takahashi K, Case BW, Dufresne A, Fraser R, Higashi T, Siemiatycki J (1994) Relation between lung asbestos fibre burden and exposure indices based on job history. Occup Environ Med 51:461–469

Thurston SW, Eisen EA, Schwartz J (2002) Smoothing in survival models: an application to workers exposed to metalworking fluids. Epidemiology 13:685–692

Tuyns AJ, Esteve J, Raymond L, Berrino F, Benhamou E, Blanchet F, Boffetta P, Crosignani P, Del Moral A, Lehmann W, Merletti F, Péquignot G, Riboli E, Sancho-Garnier H, Terracini G, Zubiri A, Zubiri L (1988) Cancer of the larynx/hypopharynx, tobacco and alcohol: IARC international case-control study in Turin and Varese (Italy), Zaragoza and Navarra (Spain), Geneva (Switzerland) and Calvados (France). Int J Cancer 41:483–491

UK (1967) Carcinogenic Substances Regulations – Statutory Instrument No. 487

US Environmental Protection Agency (EPA) (2002) Non cancer health effects of diesel exhaust. In: Health assessment document for diesel engine exhaust. Prepared by the National Center for Environmental Assessment, Washington, DC, for the Office of Transportation and Air Quality, pp 5.2–5.23. EPA/600/8–90/057F

Wild P, Réfrégier M, Auburtin G, Carton B, Moulin JJ (1995) Survey of the respiratory health of the workers of a talc producing factory. Occup Environ Med 52:470–477

Zaebst DD, Clapp DE, Blade LM, Marlow DA, Steenland K, Hornung RW, Scheutzle D, Butler J (1991) Quantitative determination of trucking industry workers' exposures to diesel exhaust particles. Am Ind Hyg Assoc J 52:529–541

Zahm SH, Blair A (2003) Occupational cancer among women: where have we been and where are we going? Am J Ind Med 44:565–575

Zahm SH, Pottern LM, Lewis DR, Ward MH, White DW (1994) Inclusion of women and minorities in occupational cancer epidemiologic research. J Occup Med 36:842–847

Environmental Epidemiology

43

Lothar Kreienbrock

Contents

L. Kreienbrock (✉)
Department for Biometry, Epidemiology and Information Processing, University of Veterinary Medicine Hannover, Hannover, Germany

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_25,

43.1 Introduction

43.1.1 Issues of Environmental Epidemiology

The human environment – "the aggregate of surrounding things, conditions, or influences especially as affecting the existence or development of someone or something" (Webster's Encyclopedic Unabridged Dictionary of the English Language 1989) – is a topic of ever-increasing public awareness. Concern about the safety of the environment has stimulated controversial debates both in the general public as well as in the scientific community. Environmental safety has to meet defined standards for the protection of public health, and epidemiological knowledge has to be gathered on the impact of risk factors on human health.

Environmental epidemiology may be defined as "the study of the effect on human health of physical, biological, and chemical factors in the external environment. By examining specific populations or communities exposed to different ambient environments, environmental epidemiology seeks to clarify the relation between physical, biological, and chemical factors and human health" (NRC 1991).

Although this is a modern definition of the relation between environmental hazards and humans, the ideas of environmental epidemiology are linked to medical history. Environmental hazards were observed in ancient times. For example, Locher and Unschuld (1999) cite the antique scripts of Hippocrates, "De aere aquis locis" or Aristotle, recommending that cities have to be located in a healthy environment and that air and water should be clean so as not to impair human well-being.

Another example of an environmental issue is the radon problem, which was first addressed in 1492 by Paulus Niavis in his essay, "Ludicium Iovis" oder *Das Gericht der Götter über den Bergbau*:

> Wie man vom Schneeberge und von den Gruben zu sprechen hat: Die arbeiten darin, und die Luft im Berge, die sehr ungesund ist, nimmt ihnen die natürliche Farbe, sehr oft geschieht es auch, dass sie frühzeitig mit Tod abgehen. (cited by Schüttmann 1992)

More than 500 years ago, the very first observations were made of a possible relationship between "the air within the mine" (*die Luft im Berge*) and symptoms of disease and early mortality in the area of ancient ore mining around the village of Schneeberg in Saxony, Germany. Writings by famous early modern physicians like Agricola, *Bermannus oder über den Bergbau* (1555) and *De Re Metallica* (1556), or Paracelsus, *Von der Bergsucht und anderen Bergkrankheiten* (1567), report cases of a special disease and describe disease clusters in Schneeberg, in Joachimsthal, and in other villages in this area. The disease was thus called Schneeberg lung disease (*Schneeberger Lungenkrankheit*) (Schüttmann 1992).

Agricola and Paracelsus did not know the real cause of the disease; it was more than 300 years until the *Schneeberger Lungenkrankheit* was identified as lung cancer in the year 1879. With this the occupational causality was stated, but the real source of exposure was not known until the processes of radiation were discovered at the beginning of the twentieth century. In 1909, the mines around Schneeberg were investigated and radioactivity was measured. Based on this, a director of the mines formulated in 1913 the hypothesis of an exposure-effect relationship between the

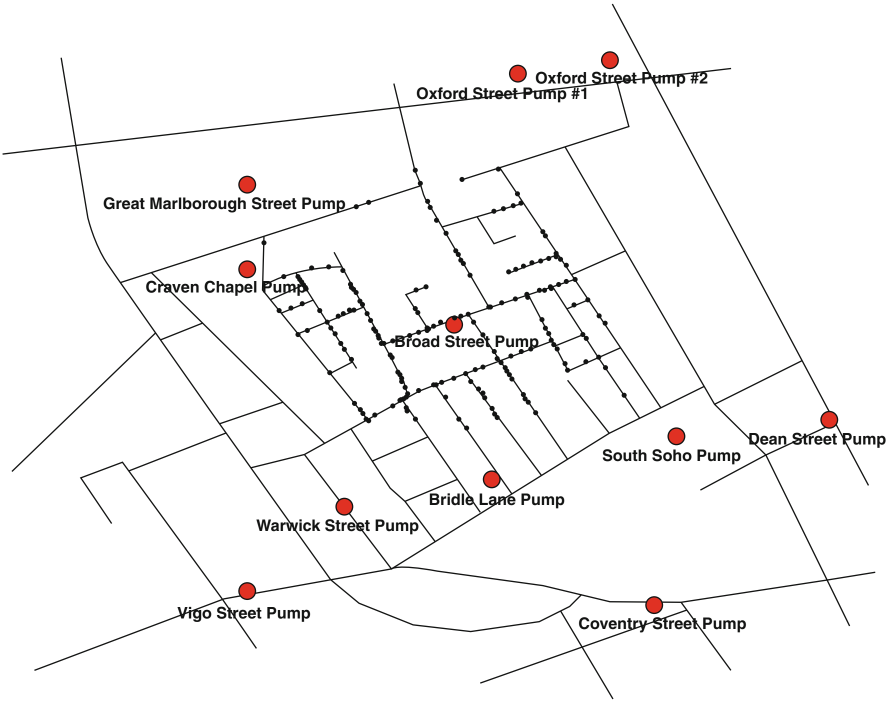


Fig. 43.1 John Snow's map of cholera cases in Soho, London (Figure modified from EpiInfo^{TM} 2000)

"radium emanation" or radon and lung cancer. Although the real nature of the dose-effect relationship was not known, the *Schneeberger Lungenkrankheit* was officially recognized in 1925 as an occupational disease in Germany in cases of diseased workers that had been extensively exposed in the mines. These observations led to the question of whether radon is a common risk. Due to the ubiquitous presence of the gas, it can be extrapolated that radon constitutes an environmental hazard in the general population (Schüttmann 1992).

Another classical example of environmental epidemiology is the famous risk map of John Snow, who reported cases of cholera in Soho, London, in the middle of the nineteenth century (Fig. 43.1).

The increasing number of cholera cases in London in 1849 and 1853/1854 caused a great awareness of the real reasons for the epidemic. John Snow's map can be considered as one of the initial steps in spatial statistics and spatial epidemiology (cf. chapter ▶Geographical Epidemiology of this handbook). It identified the source of the epidemic in the contaminated water of the Broad Street Pump. Additional research was carried out, and it was discovered that cholera incidence differed substantially in the water supply areas of the Lambeth water company and the Southwark and Vauxhall water companies. While in "Lambeth's area," 461 cases were observed in a population of 173,748, 4,093 cases occurred in a population of

266,516 in the “Southwark and Vauxhall area.” The risk ratio of 5.8 indicated a serious problem with water contamination in the “Southwark and Vauxhall area.”

These classical examples illustrate that the impact of the environment on human health was realized very early and that this impact depends on the social and political situation of a population. Nowadays there are different environmental health concerns in the developed and the developing countries. With the foundation of modern epidemiology by Sir Richard Doll and others in the middle of the twentieth century, the general focus of public and scientific concern was first on risk factors like smoking, occupation, and nutrition and their association with cancer and cardiovascular diseases. Many methods of modern epidemiology were developed as applications on these associations.

In the developed world, environmental issues of public and scientific concern are ambient air pollution, radiation (from the general environment as well as from nuclear power plants and facilities), environmental tobacco smoke (ETS), or special agents known as hazardous from occupational exposures like lead and mercury and their impact on human health. In contrast, in the developing world, the major concerns are clean air, sanitation, and clean water. Infectious diseases (cf. chapter ▸Infectious Disease Epidemiology of this handbook) in particular are a part of this problem as well as severe environmental pollution and exposures to chemical agents.

Incomplete understanding of causes of many common diseases in both developed and developing countries focuses interest on identifying environmental hazards and incorporating this knowledge in strategies of risk management.

43.1.2 Concepts of Environmental Epidemiology and Toxicology

The concept of common risk analysis has been suggested as an approach to gaining basic comprehension of all processes necessary to understand and manage risks on human health (cf., e.g., Graham 1997). Risk analysis could be defined as a three-stage process including risk assessment, risk management, and risk communication. Risk assessment may be understood as a purely scientific process, in which information is collected to describe risk factors and their impact on human health, while risk management and risk communication describe the political and social process to put this information into population-relevant actions, rules, and laws.

The general principles of a scientific environmental risk assessment may be described as a four-step process:

- Hazard identification, that is, does the agent have the potential to cause an adverse effect?
- Exposure assessment, that is, what exposures are experienced or anticipated under relevant conditions?
- Exposure-outcome assessment, that is, what is the relation between exposure and outcome in humans?
- Risk characterization, that is, what is the estimated incidence and severity of an adverse effect in a given population?

The basic scientific methodologies to deal with environmental risks and to answer the questions of risk assessment are environmental epidemiology and environmental toxicology. Both concepts should be recognized as scientific partners within a common risk assessment.

The risk assessment process in environmental toxicology usually comprises several steps of in vitro and animal experiments over a broad range of exposure intensities, leading to a NOEL (no observed effect level) in a sensitive animal species and the ADI (acceptable daily intake) that incorporates an appropriate safety factor, usually 10, 100, or higher. This process of risk assessment to protect humans works well for numerous compounds, but it also has a number of shortcomings.

First, many agents like cancerogenic, genotoxic, and allergenic chemicals are not considered to have a threshold exposure level and may induce cancer or an allergic reaction at extremely low exposures. Second, usually single compounds are tested, although we are exposed to numerous agents simultaneously. The possibility of both synergistic and antagonistic interactions between agents greatly complicates the risk assessment process. Of particular importance are synergistic reactions that have been demonstrated in vitro and in experimental animals in some instances. Here, a single agent does not induce any – or only minor – adverse effects, while the simultaneous administration of two agents elicits a strong toxicological effect. Obviously, the NOEL-ADI approach using single-agent administration does not incorporate the possibility of agent interactions. Therefore, additional methods must be developed to approach this problem. Third, the current concept does not take previous exposure into account. That is of special relevance for persistent bioaccumulative compounds like heavy metals. For these compounds, the ADI should be based on estimation of human body burdens and body burden-response relationships.

Another problem is the difficulty of extrapolation to very low doses. The restriction in the number of animals per group that can be used and the paucity of data on the mechanistic action of agents are the main reasons for these difficulties. Much effort is being spent, and more should be on the development of alternative methods in risk assessment to reduce the use of experimental animals. A number of cell and organ cultures of animal and human origin have been developed. The relative simplicity of these systems in comparison to whole animal models can be both advantageous and disadvantageous. However, few in vitro systems available at present are accepted as alternatives to whole animal models both in the scientific community and in regulatory agencies. It is clear that in vitro systems cannot reflect the entire organism and thus can reduce and refine, but not entirely replace, animal experiments. However, in vitro systems can be of enormous help to understand the mechanisms of toxic action.

These limitations of in vitro and animal experiments in toxicology call for human data from observational epidemiological studies. In general, epidemiological investigations of environmental hazards have to be conducted with the same care as in other fields, but in environmental studies, some special features have to be considered.

On the one hand, an environmental risk factor may cause severe diseases such as cancer or strong respiratory symptoms; on the other hand, these outcomes are also induced by many other risk factors; therefore, the relative impact of the environmental agent is small. This is true for many environmental problems, for example, the effects of air pollution on respiratory health, the effects of water contaminated with heavy metals and subsequent permanent damage to children's intellectual potential, and the effects of residues in food on human health. Furthermore, there are only a few situations in which high concentrations of an environmental agent affect a large part of the population. These cases may be related to "special sources" and often result from an accident like the Chernobyl disaster or from exposure of a population within a restricted area as a consequence of chronic pollution of the soil, for example, by an industrial plant.

Investigation of small risks is one of the main characteristics of environmental epidemiology. At this point it has to be noted that a small relative risk does not mean that the risk is not important. Many environmental risk factors are ubiquitous, and large proportions of a population can be exposed. This may cause population-attributable risks that cannot be neglected and may therefore be an important task for risk management. The dimension of this problem is demonstrated in several reports of the (environmental) burden of disease (cf., e.g., Prüss-Üstün et al. 2003). In assessing the health risk of environmental tobacco smoke (ETS), Heidrich et al. (2007) calculated the attributable risk (*AR*) for coronary heart disease and Heuschmann et al. (2007) calculated the *AR* for stroke in the entire German population (approx. 80 million). Although the association of ETS with both diseases is weak, that is, the corresponding relative risk is in the order of 1.2, ETS causes more than 2,000 additional deaths due to coronary heart disease and more than 700 deaths due to stroke per year.

But addressing small risks for multifactorial diseases leads to several problems in conducting an epidemiological investigation. First, all possible risk factors linked to a disease outcome should be incorporated into a study to avoid bias due to confounding and to study the possible interactions among all risk factors. This in turn requires that a large number of parameters be incorporated into a risk model. Therefore, large sample sizes are necessary to provide sufficient statistical power. Besides financial constraints, the logistic requirements are a major issue of the fieldwork in an environmental epidemiological investigation.

Second, the overall power of an investigation will be influenced by the definition of exposure itself. For example, "air pollution" is a mixture of hundreds of agents. If for this purpose an exposure-disease relationship has to be detected, the definition of exposure has to be clarified in advance. There are many different possibilities to assess the exposure and to conduct measurements, and each of them may contribute to the real exposure-disease relationship. As a consequence, a proper exposure assessment is of substantial importance, and much effort has to be made to conduct a powerful investigation.

Therefore, special methodological issues have to be taken into account in environmental epidemiology. Before these are described, special research situations need to be addressed in more detail as typical examples.

43.2 Examples of Research Fields in Environmental Epidemiology

There are numerous exposures and disease outcomes in environmental epidemiology, and there are many bibliographies, dictionaries, and structured databases with detailed information on the risks of environmental exposures. Only a few of them can be discussed further here.

43.2.1 Outdoor Air Pollution

From the earliest times, the impact of air pollution has been a major public health issue of interest in environmental epidemiology. For example, as early as 1294, the mayor of Venice is reported to have given orders to manufacturers of metals like mercury or tin to change the company's location to avoid exposing the population to "unhealthy smoke" (Locher and Unschuld 1999).

Industrial air pollution, coal burnt in domestic hearths, and traffic in combination with special weather conditions were the reasons for the London smog episode in early December, 1952. Smoke and sulfur dioxide pollution increased dramatically during this time. The visibility in central London dropped to a few meters, and there was an up to ten-fold increase of ambient air concentrations of sulfur dioxide. Traffic and general public life were strongly restricted. More than 4,000 deaths were attributed to the air pollution during that time, and mortality increased upon the average even in the months after this environmental disaster.

The London smog episode may be considered the beginning of quantitative risk assessment of the effects of air pollution on human health; it continues to influence the ideas and methods of environmental epidemiology today, as is shown, for example, by reanalysis of the 1952 data (e.g., Bell et al. 2004). As a political and social consequence, the British government introduced its first "Clean Air Act" in 1956 to reduce air pollution. As a scientific consequence air pollution was first defined by the level of sulfur dioxide and the concentration of particulate matter. Afterward, it was possible to make a more detailed analysis of the relationship between air pollution and human health by looking more closely at the diameter of particles, nitric (di)oxide, ozone, and other constituents.

The outcome of concern during the London smog was daily mortality, which was approximately four times above the average daily mortality during December. The deaths attributed to the smog were primarily linked to pneumonia, bronchitis, tuberculosis, and cardiovascular failures. This caused a controversy on whether these deaths occurred only among people with previous severe illness or whether the smog caused additional deaths. Another problem is the aggregation of cause-specific mortality rates by day. In such aggregated data, there may be perfect correlation between the exposure in question and other health risks like smoking which might confound the association. Individual-based studies are better suited for the investigation of multifactorial diseases, because perfect correlation between variables is very unlikely and it is thus feasible to separate their effects.

Since then, many studies have been conducted to resolve this debate. One of the most famous is the "Six City Study" conducted as follow-up study by the Harvard School of Public Health in the US cities Watertown, Massachusetts; Portage, Wisconsin; Topeka, Kansas; Kingston/Harriman, Tennessee; St. Louis, Missouri; and Steubenville, Ohio. In this concurrent cohort study with a 14- to 16-year mortality follow-up, the effects of air pollution on mortality were estimated, and individual risk factors were monitored for 8,111 adults. The Harvard group found that higher levels of fine particles and sulfate were associated with an increase in mortality by 26% (95% confidence interval: 8–47%), when the most polluted city was compared to the least polluted city. A positive association of mortality with concentrations of fine particles was found for cardiopulmonary disease. The authors therefore concluded: "Although the effects of other, unmeasured risk factors cannot be excluded with certainty, these results suggest that fine-particulate air pollution, or a more complex pollution mixture associated with fine particulate matter, contributes to excess mortality in certain U.S. cities" (Dockery et al. 1993).

Other studies corroborated the finding that long-term residence in cities with elevated ambient levels of air pollution is associated with an increase in mortality (e.g., Pope et al. 1995; Anderson et al. 1996; Katsouyanni et al. 1997; Samet et al. 2000; Vedal et al. 2003), but many questions still remain open. For example, a reanalysis of the "Six City Study" suggested that there is a modifying effect of education on the relationship between air quality and mortality (Krewski et al. 2004).

The epidemiology of air pollution is a highly complex and grave public health issue: On the one hand, mortality is only the endpoint of one possible health outcome; respiratory and cardiovascular disease or cancer are other serious issues. On the other hand, additional agents may be identified as characterizing parts of air pollution. Not surprisingly, the number of epidemiological studies of the impact of air pollution on human health is overwhelming (cf. WHO 2000; Brunekreef and Holgate 2002).

43.2.2 Residential Radon

A major issue of epidemiological research during the last decades has been the development of exposure-risk functions between radiation and different (cancer) outcomes. The main sources of the average annual radiation exposure in industrialized countries are medical examinations and therapies, inhalation of radon and its progeny, ingestion of natural radiation sources, and cosmic and terrestrial radiation (UNSCEAR 2000). Other sources, like the fallout from nuclear weapon experiments, the Chernobyl disaster, or the occupational exposure of workers in the nuclear industry, are low or affect only a very small part of the general population.

Worldwide the average annual effective dose from radon and its decay products is estimated at 1.15 millisievert (UNSCEAR 2000). Therefore, natural radon seems to be an environmental risk factor of great interest. Its harmful character was recognized very soon after the discovery of radiation at the beginning of the

Table 43.1 Individual and pooled results of 11 cohort studies in miners on radon exposure and lung cancer (Lubin et al. 1994)

Study	# cases	(ERR/WLM)%	CI
China	980	0.16	0.1–0.2
Czechoslovakia	661	0.34	0.2–0.6
Colorado	294	0.42	0.3–0.7
Ontario	291	0.89	0.5–1.5
Newfoundland	118	0.76	0.4–1.3
Sweden	79	0.95	0.1–4.1
New Mexico	69	1.72	0.6–6.7
Beaverlodge	65	2.21	0.9–5.6
Port Radium	57	0.19	0.1–0.6
Radium Hill	54	5.06	1.0–12.2
France	45	0.36	0.0–1.3
Pooled	*2,701*	*0.49*	*0.2–1.0*

ERR excess relative risk, *WLM* working level month as a measure of cumulative radiation exposure, *CI* 95% confidence interval

twentieth century and its identification as the real cause of the *Schneeberger Lungenkrankheit* (cf. Sect. 43.1.1).

In the history of radiation research, many (animal) experiments and measurements have been undertaken, with the usual uncertainties in extrapolating the results to humans. However, in the middle of the twentieth century, cohort studies in (uranium) miners were started to investigate the true exposure-disease relationship between radon and lung cancer. These studies confirmed that exposure to the radioactive gas radon (^{222}Rn) and its progeny increases the risk of lung cancer among workers in the uranium and other mining industries (Lubin et al. 1994; NRC 1999; IARC 2001; Table 43.1).

The study results in Table 43.1 are reported with exposures in working level month (WLM). This cumulative measure was developed historically with the idea of there being a safe threshold for radiation exposure of workers. This was defined as 1 working level (WL) and was stated as 100 picocuries per liter, which in today's SI units is equivalent to 3.7 kilobecquerel per cubic meter (kBq m-3; 1 becquerel per cubid meter is 1 radioactive decay per second in 1 cubic meter air). With an average time of occupation of 170 hour per month, this cumulates to 1 WLM.

Overall a strong exposure-effect relationship was observed. When the occupational-related exposure scale is transformed from WLM in usual Bq·m^{-3} (for details on the assumptions and calculations of the transformation cf., e.g., NRC 1999), the overall estimate of the excess relative risk per WLM was estimated as 0.49 (Table 43.1). As 1 WLM corresponds to 170·3.7 kBq·m^{-3}, we get an estimate of 0.49/(170·3.7) = 0.08 per 100 Bq·m^{-3}, that is, per additional average exposure of 100 Bq·m^{-3}, the lung cancer risk will increase by 8%. This exposure-effect relationship is further influenced by the time since exposure (TSE), the age at exposure, and other risk factors like the exposure to arsenic or other dusts and the smoking habits of the workers. These factors were not investigated deeply within the cohorts, and therefore uncertainties remain about the dose-response relationship.

The public health concern about major radioactive exposure, however, is the long-term exposure of the general population to the much smaller concentrations of radon in homes. When the exposure-effect relationship from the miner studies is extrapolated to the level of radon exposure in homes, an average population-attributable risk of lung cancer of 5–15% can be estimated in industrialized countries (Lubin and Steindorf 1995; Steindorf et al. 1995; Darby et al. 2001; Menzler et al. 2008). If this extrapolation is true, exposure to residential radon is the most hazardous environmental risk factor for cancer in the general population, and risk management strategies need to be developed.

However, a direct transfer of the risk estimates derived from miner studies or animal experiments to residential environments may not be appropriate due to substantial differences in the levels of radon exposure; other physical factors such as breathing rate; the distribution of aerosol particles size; the unattached fraction of radon progeny; confounding factors such as smoking, asbestos, and other occupational risks; nutrition; leisure time activities; social conditions; genetic susceptibility; and age, gender, and regional circumstances.

In the past 20 years, a series of well-conducted epidemiological case-control studies investigated the risk of lung cancer in relation to indoor radon exposure in the general population (Schoenberg et al. 1990; Blot et al. 1990; Pershagen et al. 1992, 1994; Létourneau et al. 1994; Alavanja et al. 1994, 1999; Ruosteenoja et al. 1996; Auvinen et al. 1996; Darby et al. 1998; Field et al. 2000; Kreienbrock et al. 2001; Lagarde et al. 2001; Tomášek et al. 2001; Oberaigner et al. 2002; Wang et al. 2002; Barros-Dios et al. 2002; Kreuzer et al. 2003). Two collaborative analyses of these studies (Darby et al. 2005; Krewski et al. 2006) supported the hypothesis of an association, despite some heterogeneity among single studies, and concluded that residential radon exposure increases the risk of lung cancer, particularly in smokers, and recent ex-smokers.

43.2.3 Non-Ionizing Radiation

Exposure to electric and magnetic fields (EMF) in the occupational as well as in the residential environment is a matter of public concern in the developed world. The debate on health consequences is widespread, and there is speculation on possible effects of exposure ranging from the development of cancer to headaches, depressions, or general malaise.

A first investigation of Wertheimer and Leeper (1979) brought EMF and childhood leukemia to the special attention of the public. The authors conducted a case-control study with 155 cases and the same number of controls, and found a risk ratio of around three. So-called wire codes of distances to electric power lines were introduced as a measure of exposure to EMF.

Further studies followed. Greenland et al. (2000) and Ahlbom et al. (2000) summarized the results of these studies by pooling them into two joint analyses in the USA and in Europe (Fig. 43.2).

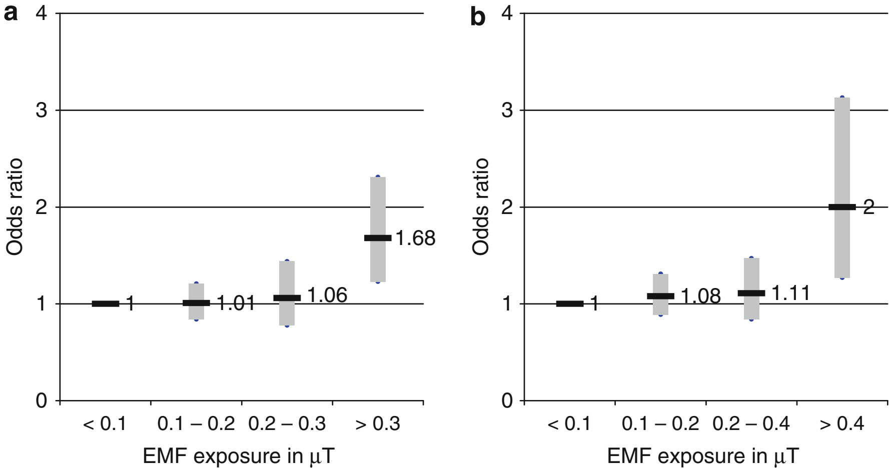


Fig. 43.2 Leukemia risk in children due to EMF exposure: odds ratios and 95% confidence intervals by EMF exposure in μT from pooling studies (**a**) in the USA (Greenland et al. 2000) and (**b**) in Europe (Ahlbom et al. 2000)

Figure 43.2 suggests a similar risk pattern in the USA and in Europe, with an overall statistical significance in the highest exposure group. However, the discussion on EMF risks is ongoing, primarily due to limitations of exposure assessment for the use of wire codes as well as for recent techniques of (personal) measurements and exposure assessment.

Habash et al. (2003a) therefore stated: "Currently, the evidence in support of an association between EMF and childhood cancer is limited, although this issue warrants further investigation. Evidence of an association between EMF exposure and adult cancers, derived largely from occupational settings, is inconsistent, precluding clear conclusions. There is little evidence of an association between EMF and non-cancer health effects. . . . Further research is needed to clarify the ambiguous findings from present studies and to determine if EMF exposure poses a health risk."

43.3 Special Methodological Issues in Environmental Epidemiology

Although there is great variation in study designs and in measurement techniques, methods of exposure assessment and statistical analyses used in environmental epidemiology share common features that will be summarized in the following.

43.3.1 Principles of Study Design

As mentioned above, one of the major problems in studying an environmental risk factor is often the low risk associated with it. Therefore, the overall power of a study has to be large to increase the chance of identifying health hazards. The proper choice of the study design is thus a basic step in conducting an investigation in environmental epidemiology.

43.3.1.1 Standardization of Study Designs

The power of an epidemiological study is related to two major components, namely, the sample size and the precision of all instruments used. The number of available cases and the financial budget are constraints that determine the limitations of every investigation. Therefore, it is desirable to standardize study designs to increase the precision of the results and to establish a basis for comparisons between different studies.

For example, the designs of studies of the impact of residential radon developed gradually over time. In a very first step, ecological studies were performed to investigate the numerical correlation between radon concentration and lung cancer mortality with aggregated data, for example, on the basis of administrative districts (Stidley and Samet 1993). This type of study design is very popular in environmental epidemiology but is strongly influenced by different types of biases, especially confounding bias (cf. chapter ▶Descriptive Studies of this handbook). Figure 43.3 shows an example of an ecological analysis for radon data in West Germany.

Figure 43.3 shows a typical problem which occurs in ecological studies in environmental epidemiology when the environmental risk factor is associated with a low risk and the disease is influenced by a strong risk factor that is itself associated with the risk factor under study. On the one hand, in Germany – as in other industrialized countries – smoking is much more popular in cities and urban areas (e.g., the major cities Berlin, Hamburg, Düsseldorf, Bremen) than in rural areas (e.g., the rural districts Niederbayern, Oberpfalz, Tübingen in the south

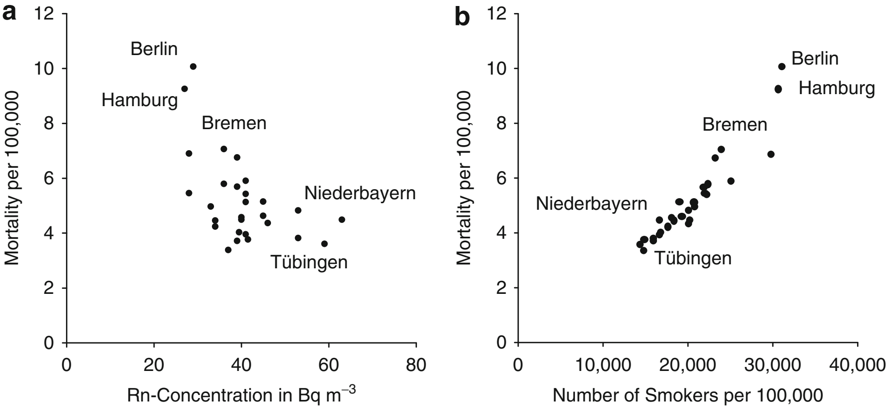


Fig. 43.3 Lung cancer and radon in West Germany. (**a**) Lung cancer mortality in women by average radon concentration in West German districts (*Regierungsbezirke*). (**b**) Lung cancer mortality in women by average smoking prevalence among females in West German districts (Kreienbrock et al. 2012, modified)

Table 43.2 Three generations of radon studies with individuals (Pershagen 1993, personal communication)

Generation	Characterization
I	– Conducted in the early 1980s – Small sample sizes – No or crude Rn measurements – No or only small control for confounding
II	– Conducted in the late 1980s – Small sample sizes – Rn measurements in one house – Some control for confounding
III	– Conducted in the 1990s – Big sample sizes – Rn measurements in all homes – Control for confounding

of Germany). In addition, smoking prevalences decrease from north to south. On the other hand, radon concentration is low in cities and urban areas, while higher measurements are much more likely in rural areas. This pattern causes a "protective radon effect" if the smoking habits are neglected.

This type of misleading findings in ecological studies is known as the ecological fallacy and can be observed in many studies on radon (Cohen 1993, 1997) as well as on other environmental risk factors. Therefore, ecological studies are only reasonable as a starting point for discussion. Causal relationships could not be concluded from this type of study design.

Study designs based on individuals are necessary to investigate the causal role of environmental risk factors. For studies on the radon-related lung cancer risk in the general population, Pershagen (1993, personal communication) described three generations of studies following the ecological studies (see Table 43.2).

The studies from the first and second generations were very small and therefore did not have the power to detect the radon-related lung cancer risk, which is expected to be low. Thus, in 1989, 1991, and 1993, workshops were organized with all principal investigators of ongoing studies during that time to develop "Guidelines for Conducting Radon Studies" (Samet et al. 1991). These workshops led to a standardization of study protocols and formed a basis for future pooling of studies.

Although this process of standardization cannot be established as a general rule in environmental epidemiology, it should be recognized as a strategy that may enhance the overall power of investigations.

Because the risk ratio of an environmental risk factor is usually low and many of the diseases under study are not frequent, a case-control study is the type of study design that is most appropriate to investigate an environmental risk factor within the common population. This is usually done for studies on radon and lung cancer as well as for most of the studies investigating the impact of environmental agents on cancer. For example, of the 20 studies on residential radon mentioned in Sect. 43.2.2, only the Czech study was conducted as a cohort study in which a case-control study is nested (Tomášek et al. 2001).

In contrast, at first glance cohort studies seem to be rare in environmental epidemiology due to the extensive efforts they require. Nevertheless, there are well-known examples of this study design, such as the "Six City Study" on air pollution and mortality (Dockery et al. 1993, cf. Sect. 43.2.1), and other longitudinal designs, studying long-term effects of the relation between air pollution exposure and public health (Brunekreef 2003). Therefore, the use of cohorts is increasing in different fields of environmental interest. For example, cohorts are now being used in investigations on the impact of nutrition on human health, like the EPIC project (European Prospective Investigation into Cancer and Nutrition), which studies more than 500,000 participants in ten European countries (Slimani et al. 2002). However, the establishment of large cohorts for environmental research is often facing financial restrictions limiting the proper assessment of exposure (see below).

43.3.1.2 Selection of Study Participants and Choice of the Study Area

The first step in planning an investigation in environmental epidemiology is the choice of the study area. For this purpose, monitoring data of measurements is very helpful to find out whether the risk factor is present within a defined region and which variation it shows. This is necessary to distinguish between exposed and non-exposed populations or to evaluate an exposure-effect relationship for a common range of possible exposures.

However, areas with high average concentrations or with a greater proportion of elevated values do not by definition guarantee a proper environmental study. If the risk factor is distributed equally, cases and controls of the study region will be exposed to a similar degree, and no effect of the risk factor can be observed in the study. This problem is known as "overmatching on the exposure factor." It can be avoided by separating areas with different levels of the risk factor. This problem is commonly discussed in studies on the possible impact of outdoor air pollution as a risk factor (e.g., Dockery et al. 1993; Pope et al. 1995; Heinrich et al. 2000; Hoek et al. 2000).

A different problem occurs in studies on the indoor environment: The concentration in homes is influenced by outdoor air, geology, or other factors related to the outside environment but will be modified by the type of house and the (ventilation) habits of the inhabitants. This may result in low measurements even in areas with a great impact of outdoor air or geology on increased concentrations of the agent under study. Taking into account the fact that measurements are only one component of the exposure assessment (see below), it is possible that both cases and controls will be exposed at a very low level. This effect was observed, for example, in many radon studies, for instance, in the study in Winnipeg, Canada (Létourneau et al. 1994), or the study in West Germany (Kreienbrock et al. 2001), where measurement programs showed areas with high radon concentrations on average, while exposures of study participants were low. This was mainly due to the mobility of the population that resulted in low average exposures of individuals (Warner et al. 1996).

Therefore, the choice of the study area has to be influenced both by the presence of a monitoring program indicating areas with elevated concentrations of the risk factor and by a population density that yields a sufficient number of study participants.

The selection of cases within the study area is dependent on the health system within this area. If national or regional registers are available, cases will usually be recruited from these registers; if not, case recruitment from hospitals will be the usual sampling strategy. The problem of a proper selection of a group of cases is similar in all kinds of epidemiological applications, but if measurements on environmental agents have to be conducted, this causes additional problems, especially if the cases are to be recruited in hospitals and the measurement campaign is (technically) complicated. This will decrease the response proportion in the case group, and the rates may be differential by disease stage and thus may cause a selection bias. On the other hand, even in studies with registers, this type of problem is present, especially if next-of-kin interviews have to be conducted instead of interviews of cases.

Similar problems due to selection procedures and exposure assessment may be present for the control group in environmental studies. As in any other epidemiological study, the different types of bias may occur, but these may be strongly influenced by the type of measurement to be conducted. This problem can be observed mainly in "special circumstances studies," where recent environmental problems of great public interest are studied. This may cause different response proportions among subgroups within the general population, which may introduce a severe selection bias and impair comparability between subgroups.

An example of a special circumstances situation was described by Oberaigner et al. (2002) for radon epidemiology. In 1989, high lung cancer mortality rates were reported for the District of Imst, Tyrol, Austria, in an alpine region without any industry. First investigations in this district (Ennemoser et al. 1994a, b) identified one village with extremely high radon concentrations; the highest concentrations measured in residences were above 100,000 Bq·m^{-3}. In the meantime, detailed cross-sectional investigations have examined both the radon gas concentrations and the geology. In fact, only half of the village was affected, and about 40 residences were identified with concentrations above 1,000 Bq·m^{-3}. There were only few and isolated high concentrations in the other villages of this region. In fact, the scientific interest in this area was linked to population's interest and response. While in the beginning the whole population was interested, in cross-sectional measurement programs, there was a gradual decrease in response proportions linked, unfortunately, with increasing exposure.

Decreasing response proportions were also observed during the 1990s after the reunification of Germany, when many environmental measurement programs identified numerous areas with elevated levels of environmental agents of scientific interest in the former German Democratic Republic (Heinrich et al. 2002; Frye et al. 2001). One may call this effect a possible bias due to "over-examination."

These examples indicate that selection bias is of special interest in environmental epidemiology and special efforts have to be undertaken during fieldwork to avoid possible selection effects. Besides an exact definition of inclusion and exclusion criteria of study participants, selection effects are mainly dependent on a proper documentation of the reasons for response and non-response combined with an additional investigation on the non-responders. Moreover, it is necessary to make corrections for selection bias.

43.3.1.3 Sampling Procedures and Correction for Non-Response Bias

Usually the recruitment of study participants is described as a process of selecting a study population from a target population by means of a defined sampling scheme. If the study population is a representative sample of the target population, the study results should be unbiased; otherwise, a selection bias may occur which will have to be discussed. For studies in environmental epidemiology, these biases may be described from two different perspectives.

On the one hand, different selection procedures for the exposed and the non-exposed populations are possible, as observed in many environmental studies. Heinrich et al. (2002) described studies on air pollution and respiratory health and allergies in the former German Democratic Republic, where response proportion decreased with increasing exposure. On the other hand, response proportions may differ between cases and controls of a study. This effect is often due to the fact that controls are less motivated so that their response proportions are lower than that of the cases who want to know whether environmental hazards are responsible for their disease.

Investigation of the non-response patterns seems to be necessary to investigate these processes and to establish a basis for an adjustment for possible selection effects caused by non-response. The general principle of a non-response investigation was initially outlined by Hansen and Hurwitz (1946) and is illustrated in Fig. 43.4. It may be assumed that study response is a characteristic of the members of the target population that partitions the population in two strata, N_1 responders and N_2 non-responders. An epidemiological study of study size n may be interpreted as a stratified random sample from the target population, that is, a stratified sample of real responders n_1 and non-responders n_2. This process can be interpreted as a first phase of a sampling plan. In a second phase, a real sample of size n_2^* is drawn from the n_2 non-responders, and all efforts are made to get information from this second-phase subsample of size n_2^*, for example, by conducting additional telephone interviews. The final study population n^* will be calculated as the sum of these subsample sizes.

If a non-responder investigation is conducted, adjustment for non-response will be possible in a straightforward way. For simple linear statistics like prevalences or means (of exposures), an adjusted estimator can be computed as a simple weighted mean of the two strata means of responders, denoted as $\bar{y}_1$, and non-responders, say $\bar{y}_2$, that is,

$$\bar{y}_{\text{adj.}} = \frac{n_1}{n^*} \cdot \bar{y}_1 + \frac{n_2^*}{n^*} \cdot \bar{y}_2. \tag{43.1}$$

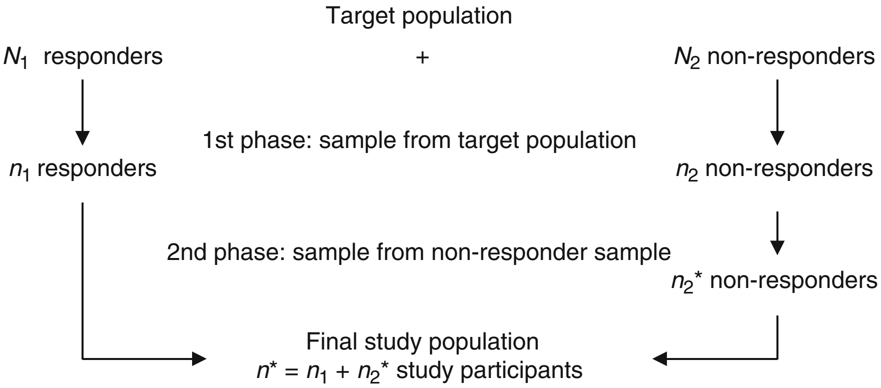


Fig. 43.4 Two-phase sampling scheme for collecting sampling information on non-responders

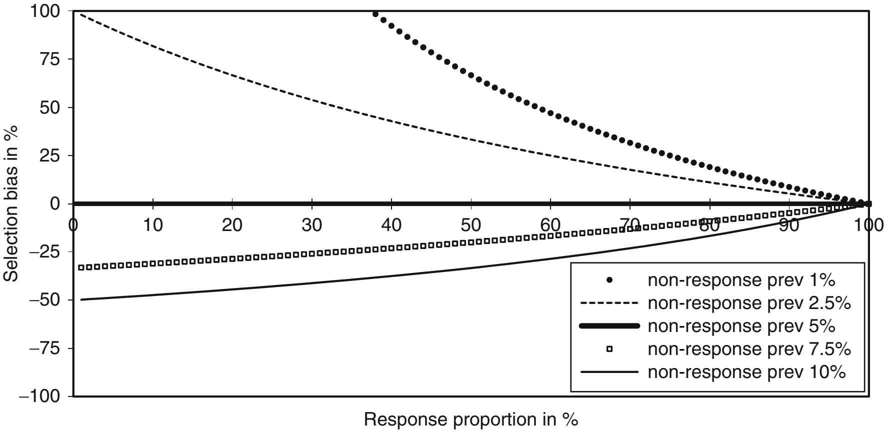


Fig. 43.5 Selection bias in % due to non-response for estimating a prevalence in the target population of 5% for varying exposure prevalences among non-responders

This type of adjustment (Eq. 43.1) is useful in environmental epidemiology, especially if diseases or exposures under study are rare and therefore prevalences are low. Figure 43.5 displays the impact of selection bias due to non-response on the estimation of a prevalence of 5% by comparing the adjusted estimator from Eq. 43.1 with the usual estimator, namely, the average $\bar{y}_1$, that is calculated without adjustment for non-response. This situation may occur, for example, in a cross-sectional study investigating the impact of air pollution on respiratory symptoms like asthma in boys. It can be shown that selection bias increases with increasing non-response, depending on the magnitude of the difference of the exposure prevalences between responders and non-responders.

Table 43.3 Realization of study subjects n_{ij} as sample with probability W_{ij} from the members N_{ij} of the target population, $i = D$ for diseased and $\bar{D}$ for non-diseased subjects, $j = E$ for exposed and $\bar{E}$ for exposed and non-exposed subjects

	Exposed	Non-exposed
Diseased	$n_{DE} = W_{DE} N_{DE}$	$n_{D\bar{E}} = W_{D\bar{E}} N_{D\bar{E}}$
Non-diseased	$n_{\bar{D}E} = W_{\bar{D}E} N_{\bar{D}E}$	$n_{\bar{D}\bar{E}} = W_{\bar{D}\bar{E}} N_{\bar{D}\bar{E}}$

A linear adjustment is not suitable for estimating ratios, and the strata of responders and non-responders have to be split up into the exposed and the non-exposed subpopulations. Kleinbaum et al. (1982) therefore introduced the 2×2 table of an epidemiological study as the selected outcome from the target population that results in a 2×2 table of selection probabilities for subjects from the entire target population (Table 43.3).

If selection of study participants is outlined as in Table 43.3, an epidemiological study to estimate a population odds ratio is unbiased if the odds ratio of the selection probabilities $W = (W_{DE} W_{\bar{D}\bar{E}})/(W_{D\bar{E}} W_{\bar{D}E})$ is equal to unity. If W is greater than unity, the study will be biased away from the null, else the bias is toward the null.

Investigations in environmental epidemiology may be very sensitive to selection bias. The odds ratio of the selection probabilities W may be used both for a quantitative and a qualitative assessment of the possible bias. If, based on the study design and the sampling techniques applied, the sampling probabilities can be computed, a quantitative adjustment for selection bias will be possible by multiplying the study odds ratio with the inverse W^{-1}.

If sampling probabilities are not available, a qualification of the direction of selection bias is possible. Examples for a qualitative assessment of the bias can be adopted from cross-sectional studies of the impact of air pollution on respiratory health. Here, the motivation of the study subjects may be influenced by the disease and by the exposure situation. In contrast, subjects who are not ill and who are not exposed may not be motivated to participate. This will decrease the selection probability $W_{\bar{D}\bar{E}}$; hence, $W < 1$, and the study odds ratio will be biased toward the null.

In contrast, if the same study is conducted in a “special circumstances area,” the effect may be vice versa if the healthy exposed population is unwilling to participate. This will decrease the selection probability $W_{\bar{D}E}$; hence, $W > 1$, and the study odds ratio will biased away from the null.

43.3.2 Measurements and Exposure Assessment

The measurement of an environmental agent such as the risk factor under study is the primary issue of an investigation in environmental epidemiology. Such measurements can be used directly or as a basis of an exposure assessment. The main problems to be addressed in any investigation in environmental epidemiology

Table 43.4 Examples of short-term and long-term exposure-effect relationships in environmental epidemiology

Exposure-effect relationship	Example
Short-term exposure/short-term effect	Traffic-related agents – acute respiratory symptoms Smog episodes – mortality
Short-term exposure/long-term effect	Fallout from nuclear weapon tests – cancer incidence Contaminated food – new variant of Creutzfeldt-Jacob disease
Long-term exposure/long-term effect	Environmental tobacco smoke – lung cancer Residential radon exposure – lung cancer

are an appropriate choice of the measurement technique, the method of exposure assessment, and the statistical evaluation to model the exposure-effect relationship.

Numerous types of exposures may be distinguished in environmental epidemiology. One useful categorization is to distinguish between short-term and long-term exposures and between short-term and long-term effects. The time interval between exposure and effect is often called time since exposure (TSE). Table 43.4 shows typical examples that may occur in environmental epidemiology within these categories.

The relationship of short-term exposures to short-term effects as well as the relationship of long-term exposures to long-term effects may be postulated as the typical situations under study in environmental epidemiology, while the situation in which a short-term (single) exposure causes a long-term effect seems to be rare. It may be argued that it is much more complicated to assess long-term exposures and that short-term exposures may be described by the measurement itself. However, this is not a general rule.

For radon and for many other agents, residential exposure is a long-term exposure starting at birth and ending with the study recruitment, for example, by the index date of diagnosis or the date of a standardized personal interview. The outcome under study is lung cancer as a long-term effect. Taking into account the fact that residential exposure is characteristic of the homes a study participant has lived in during his or her life, the overall exposure can be outlined as in Fig. 43.6.

The exposure pattern depicted in Fig. 43.6 seems to be typical for a long-term exposure to an environmental agent, that is, that individuals are exposed to different average levels with varying intensities around these levels in different time windows over their lifetimes. In the radon example, different levels of exposure are related to different dwellings the study participants have lived in during their lives. Each of the resulting time windows covers many years, a typical situation in many studies where the exposure conditions are linked to locations and these exposures are more or less constant over time. Additional examples for this situation are found in all studies investigating the impact of indoor or outdoor residential exposures on human health like the different agents responsible for air or soil pollution.

Sometimes the time windows of different exposure levels may be subdivided if there is additional knowledge about external circumstances and about the habits

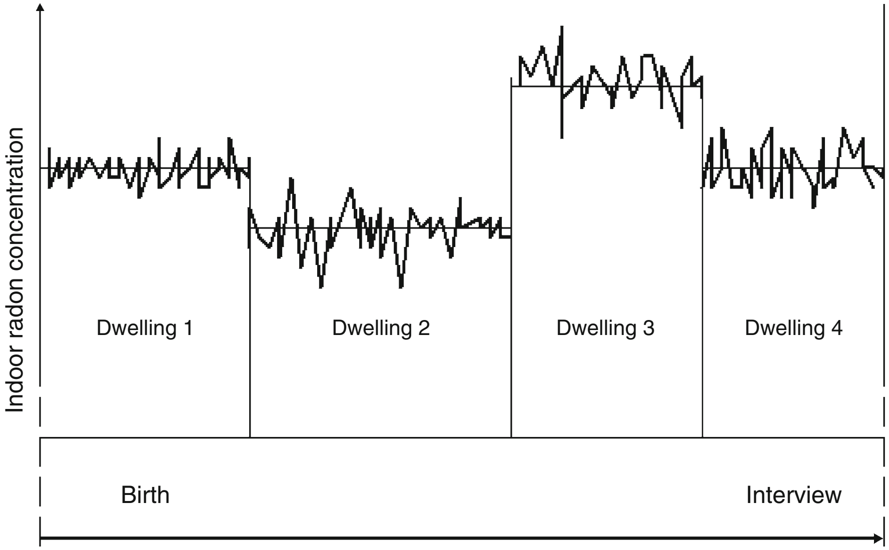
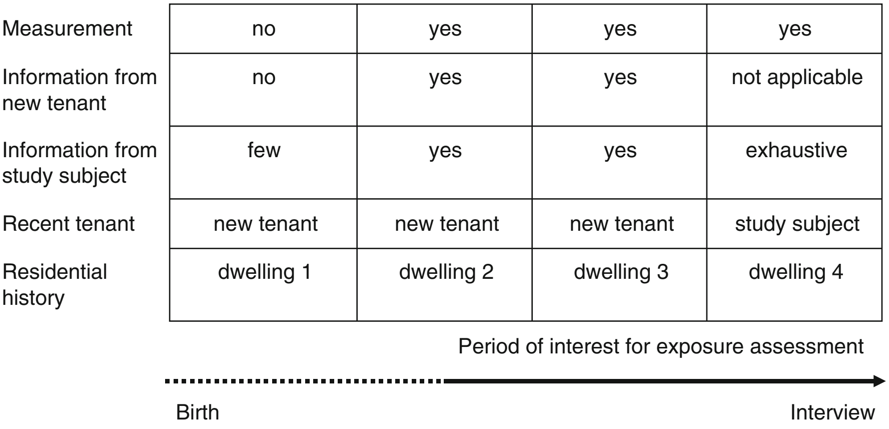


Fig. 43.6 Residential radon exposure during a participant's life (fictitious participant's biography) (Wichmann et al. 1998, modified)

of the participants. For the radon example, these modifications of exposure can be summarized as follows. It is known that radon concentrations in homes differ due to temperature and air pressure, resulting in different concentrations over the seasons (on average, the concentration in winter is twice as high as in summer) as well as over the day (higher concentrations during the night than in the daytime). From this point of view, time windows on a monthly or even on an hourly basis have to be considered. But the living habits of the participants will also influence the exposure pattern. For example, the time spent indoors during the day as well as during the year will be an important factor for an individual's exposure. In addition, exposure will be modified by living habits like ventilation behavior, by the existence and pattern of utilization of different rooms within the homes, and by other individual habits.

The combination of both dimensions of exposure windows due to the external circumstances and to the habits of the participants may provide a sophisticated structure of the exposure windows, which on the one hand will decrease the variation of exposure around an average exposure within a time window. On the other hand, the process to obtain proper information and measurements within these time windows will be much more complicated and difficult.

Different strategies to model an exposure-disease relationship are suitable in environmental epidemiology. As a very rough strategy, the real exposure may be extrapolated by means of a (categorical) variable that is used as a surrogate. Such variables and scores are widely used because no measurements have to be made and information can often be collected easily by means of a questionnaire or even by simple observation. Thus, classical categorization into exposed and



Measurement	no	yes	yes	yes
Information from new tenant	no	yes	yes	not applicable
Information from study subject	few	yes	yes	exhaustive
Recent tenant	new tenant	new tenant	new tenant	study subject
Residential history	dwelling 1	dwelling 2	dwelling 3	dwelling 4

Fig. 43.7 Information and measurement sampling scheme for a residential radon exposure (fictitious participant's biography) (Wichmann et al. 1998, modified)

non-exposed groups can be used, for example, by definition of areas with high or low industrialization or by geology. For example, early studies of the impact of radon on cancer tried to assess the exposure by categorizing the type of house as a surrogate.

Such scores are insufficient to assess an exposure pattern as in Fig. 43.6, and measurements of the agent under study should be carried out. Besides technical and financial restrictions, for epidemiological purposes it has to be clarified whether these measurements are suitable for a population-based study. Continuous measurement of the agents under study for every study subject will be rare, and the most frequently applied technique uses short-term or cumulative passive samplers. For example, for radon exposure assessment, passive sampling techniques were utilized by applying an information and measurement sampling scheme as outlined in Fig. 43.7.

The information and measurement sampling scheme given in Fig. 43.7 was used in modified versions in many studies investigating the impact of residential radon on the general population. In a personal interview with trained interviewers, the following information was recorded using a standardized questionnaire for each dwelling inhabited over a given period relevant for exposure assessment (e.g., 30 years before date of interview):

- Average time spent in each room per day, ascertained separately for each dwelling inhabited during the exposure assessment period
- Average periods of regular absence from the residence in each residence period, such as holiday; weekly absence due to occupation; etc.
- Persistent characteristics such as type of house, year of construction, type of construction, and type of basement
- Changeable characteristics such as insulation of basement and windows, heating system, and ventilation habits

- Calendar years of residence periods in each dwelling inhabited
- Calendar years of alterations of building characteristics within a residence period

This information was not collected in detail for dwellings outside the period relevant for exposure assessment. Measurements of radon concentrations were carried out by means of so-called alpha track detectors placed both in the living room and in the bedroom of the present and former dwellings of the participants. The relevant information gathered from the participants and the subsequent tenants of the former dwellings were recorded according to their periods of residence.

These data provide the basis for different methods of exposure assessment. The most popular one integrates the several measurements as a time-weighted average exposure representative of the period of exposure assessment. In addition, the exposure assessment has to adjust the current measurements for alterations of living habits and for occupancy times in the present and previous residences as well as for seasonal effects if the measurement took place in a non-typical season.

In the first approach, the exposure has to be calculated as a usual linear weighted average in the same scale in which measurements are reported. For radon concentrations, this is in $\mathrm{Bq{\cdot}m^{-3}}$. The second approach considers the cumulative exposure for a defined period before the interview. This exposure window should be the most relevant time interval with respect to the disease under study. For the lung cancer risk due to radon, time windows from 5 to 15 up to 5 to 30 years prior to diagnosis are under discussion (ICRP 1993; Lubin et al. 1994). Here, the measurements in the present homes are supplemented by measurements in the previous homes, corrected by changes due to reconstruction of the house or different ventilation habits of the study subjects and the present inhabitants and by seasonal adjustment. This cumulative exposure can be expressed in $\mathrm{Bq{\cdot}m^{-3}}$ per year.

To evaluate corrections of measurements due to changes between subjects and present inhabitants in all rooms measured (e.g., living room and bedroom), a bivariate version of a multiplicative model of the following type can be assumed:

$$rn = \mu \times \beta_1 \times \beta_2 \times \beta_3 \times \ldots \times \beta_J \times \exp(\varepsilon), \tag{43.2}$$

where rn denotes the observed radon concentration; μ is an overall baseline radon concentration; $\beta_1, \beta_2, \beta_3, \ldots \beta_J$ are J categorized effect parameters corresponding to the factors of house characteristics, ventilation habits, reconstructions, and so on; and ε is an error term from a normal distribution with zero mean.

The univariate version of Eq. 43.2 is well established in the context of radon surveys (Gunby et al. 1993; Kreienbrock and Siehl 1996). It is based both on the physical model that changing habits will lead to an exchange of a specific fraction of indoor and outdoor air and on the usual figure that indoor radon concentrations have been found to follow a log-normal distribution (Bäverstam and Swedjemark 1991; Gunby et al. 1993; Miles 1994; Lubin et al. 1995).

The log transformation of Eq. 43.2 leads to a bivariate normal distribution of the transformed radon measurements both in the living room and bedroom, which is required for standard MANOVA (multivariate analysis of variance, Anderson 1984).

An additive model can be computed which estimates final model effects $\beta_{jk}^{(i)}$, for each category k of a factor j in room i, with $k = 1, \ldots, K_j$, number of categories of factor j, $j = 1, \ldots J$, factors in the model, and $i = 1, 2$ rooms.

With $K = 1$ as the reference category for each factor j, correction factors for the measured radon concentrations of present inhabitants can be defined as

$$\text{Correction}\,{}_j^{(i)} = \frac{\exp\left(\beta_{jk_j}^{(i)}\right)}{\exp\left(\beta_{j\ell_j}^{(i)}\right)}, j = 1, \ldots, J, i = 1, 2, \tag{43.3}$$

where the participant's category is k_j and present inhabitant's category is ℓ_j relating to factor j, $j = 1, \ldots, J$ (cf. Gerken et al. 2000). These factors may be estimated from the above linear model by an ordinary least square algorithm and are then used for calculating the corrected average cumulative radon exposure per year and individual.

A correction for all measurements has to be applied to evaluate adjustments of measurements for seasonal effects. It can be assumed that the logarithm of the radon concentration in a given house follows a sine-cosine curve over 1 year. Then, the cumulative measurement obtained from the alpha track detector involves a time integral over the sine-cosine curve, that is,

$$\ln(rn_i) = \mu_i + 1/(t_{i2} - t_{i1}) \int_{ti1}^{t_{i2}} s_t \mathrm{d}t + \varepsilon_i \tag{43.4}$$

with rn_i being the observed radon concentration, μ_i the mean radon concentration in room i over the entire year, t_{i1} and t_{i2} the first and last day of measurement, ε_i an error term, and s_t the sine-cosine curve given as

$$s_t = \alpha_1 \cdot \cos(2\pi/365 \cdot t) + \alpha_2 \cdot \sin(2\pi/365 \cdot t) \text{ for } t = 1, \ldots, 365, \tag{43.5}$$

where α_1 and α_2 are parameters that can be estimated from the data within the framework of a standard linear model (Pinel et al. 1994; Oberaigner et al. 2002; Baysson et al. 2003). A result of an adjustment process like Eq. 43.4 and Eq. 43.5 is outlined in Fig. 43.8 for a radon study in Austria, where the maximum radon concentration was reached in mid-February (2.62 times the mean concentration), and the minimum concentration in mid-August (0.38 times the mean concentration).

Overall the exposure assessment may be summarized as a weighted cumulation of several stages of information based on measurements, questionnaire data, and modeling (see Fig. 43.9).

This process of exposure assessment (cf. chapter ▸Exposure Assessment of this handbook) can be considered as a very typical situation in environmental and in other fields of epidemiology. Additional examples of the calculation of cumulative exposures are the concept of working level month as a cumulative measure of the exposure to ionizing radiation among workers in the uranium mining industry (cf. Sect. 43.2.2; NRC 1999), the concept of fiber-years for the cumulative exposure

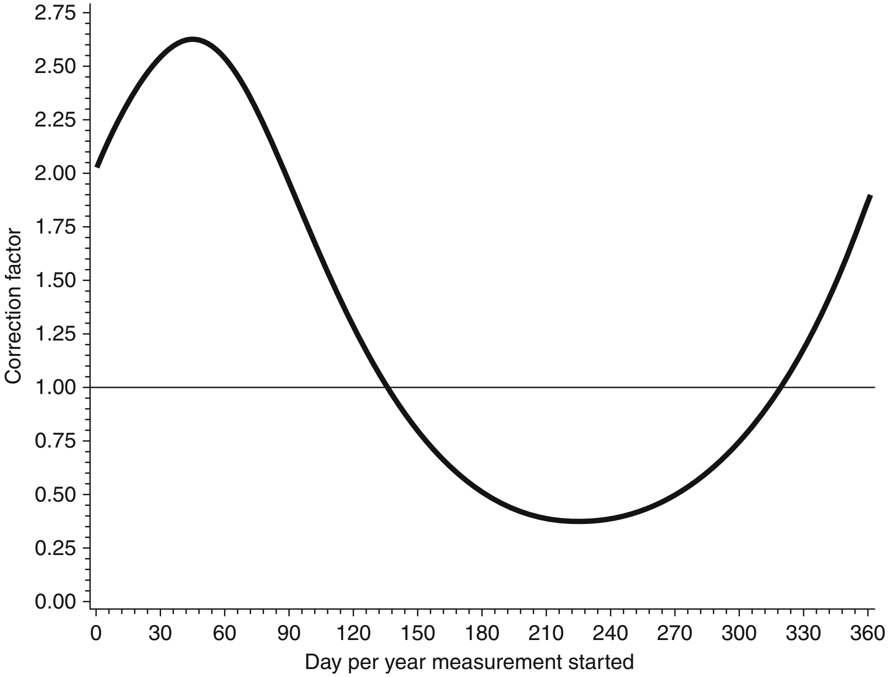
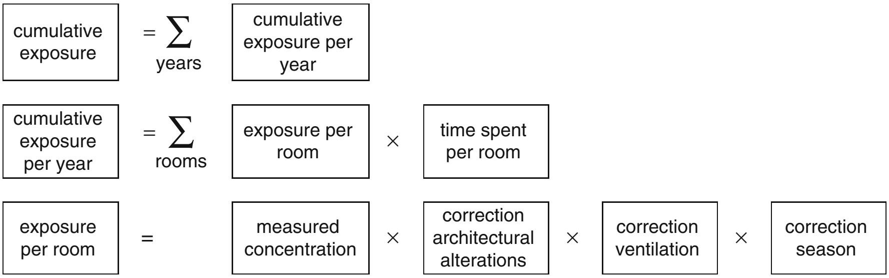


Fig. 43.8 Seasonal correction of radon measurements (Oberaigner et al. 2002, modified)


cumulative exposure = $\sum$ years cumulative exposure per year

cumulative exposure per year = $\sum$ rooms exposure per room × time spent per room

exposure per room = measured concentration × correction architectural alterations × correction ventilation × correction season

Fig. 43.9 Assessment of residential radon exposure

to asbestos, or the pack-year concept that summarizes all cigarettes smoked during the lifetime (1 pack-year = 1 pack of 20 cigarettes a day for 1 year = 7,300 cigarettes).

Although these concepts are very popular, and sophisticated solutions of integrating exposures are available, a cumulative exposure quantification may lead to substantial problems in evaluating a risk model. These strategies need retrospective information and may therefore be influenced by information bias. This is especially true if information is based on interviews and participants' memory. As a famous example, the British Doctors' Study showed that there was a large gap between the initial reporting of the smoking habits of study participants and the same subjects' answers to the same questions on his or her past habits some years later (Doll and Peto 1976).

Besides this general problem, the real nature of an exposure-risk relationship may not be well described for an agent if integrated exposures are used. It is known that the risk may differ in different exposure patterns even if cumulative exposure is constant. These effects are sometimes addressed as the inverse dose-rate effect that results in higher risks if the same cumulative exposure is reached with a low dose rate in contrast to the same cumulative exposure reached by a high dose rate over a shorter time interval. This effect was reported in cancer epidemiology for smoking as well as for exposure to asbestos or ionizing radiation. To evaluate such effects, this has to be taken into account by the risk model used.

Influenced by studies of the effect of a single exposure in a very short time interval, studies of the impact of environmental agents make use of the concept of time since exposure (TSE) that generally leads to the effect that risk decreases or even disappears if TSE is increased. This effect is able to modify the response to a cumulative exposure. If a single point in time is indicated when the an exposure occurs, TSE can be well defined. This is possible for several studies when a point in time is well defined, for example, studies of the health impact of the exposures to the atomic bombs of Hiroshima and Nagasaki, for many occupational exposures, or for exposure due to medical examinations.

However, for most of the exposures in environmental research, as for radon exposure, the concept of a single point in time is not applicable, since continuous exposures have to be taken into account (see Fig. 43.6). Therefore, additional strategies on modeling exposure have to be used.

Finkelstein's approach makes direct use of the time window structure to better understand the influence of a special risk factor over time (cf. Finkelstein 1991). This exploratory method can be described as a series of risk models that include total cumulative exposure and an additional covariate for exposure received during a fixed time interval. Characteristics of the fitted models provide insight into the influence of exposure increments on disease risk at different points in time.

Let Y_i denote the disease status of individual $i(i = 1, \ldots, n)$, and let $x_i(t)$ denote the exposure of the ith individual at time t before interview ($t \in [0, T]$), where T depends on the length of collected exposure histories. Additional covariates $z_i = (z_{1i}, \ldots, z_{mi})$ are used to adjust for confounding.

Then a time window approach sequentially fits models that include cumulative exposure to attained age, A, and cumulative exposure received over a time interval of fixed width k as covariates. Intervals of various width k can be considered. For the time window centered at time c before interview, where $c \in [k/2, T - k/2]$, a model M_c of the form

$$\text{logit}\,\Pr(Y_i = 1|z_i, ; x_i(t), t \in [0; T]) = \alpha_0 + \alpha' z_i + \beta_1 \int_0^{A_i} x_i(t)dt + \beta_2 \int_{c-k/2}^{c+k/2} x_i(t)dt \tag{43.6}$$

may be fitted, and the likelihood ratio test statistic can be computed as

$$LR_c = -2 \log \frac{\max_{\alpha,\beta} L(M_c|\beta_2 = 0)}{\max_{\alpha,\beta} L(M_c)}, \tag{43.7}$$

which compares model (43.6) with the corresponding "null" model without the time window exposure variable, that is, $\beta_2 = 0$. The value of c is then varied over its range. For fixed c, the parameter β_1 represents the increase in the log odds ratio per unit exposure, while β_2 represents the additive effect on a log scale of a unit exposure that occurred during the specific time window of length k centered at time c. The likelihood ratios between the models with and without the time window, LR_c, can be compared to assess the significance of the additional exposure variable. In this way, a continuous weighting of the impact of the exposure over time is possible (see Fig. 43.10; Hauptmann et al. 2000a, b).

Overall it may be stated that the process of sampling information and measurements to conduct an exposure assessment is the major issue in an investigation of the impact of an environmental agent on human health. Besides technical and financial constraints, several types of information bias and uncertainty influence this process, and uncertainty of exposure measurements still remains, even if detailed descriptions of the exposures are available. This should be considered during the statistical analysis of a study.

43.3.3 Statistical Analysis

As in all other fields of epidemiology, classical concepts of risk models like the categorical analyses of (stratified) contingency tables or modeling approaches like logistic regression or Cox' proportional hazards models are used in the study of the risk of environmental agents on human health to describe the relationship between exposure and disease (cf. chapter ▶Regression Methods for Epidemiological Analysis

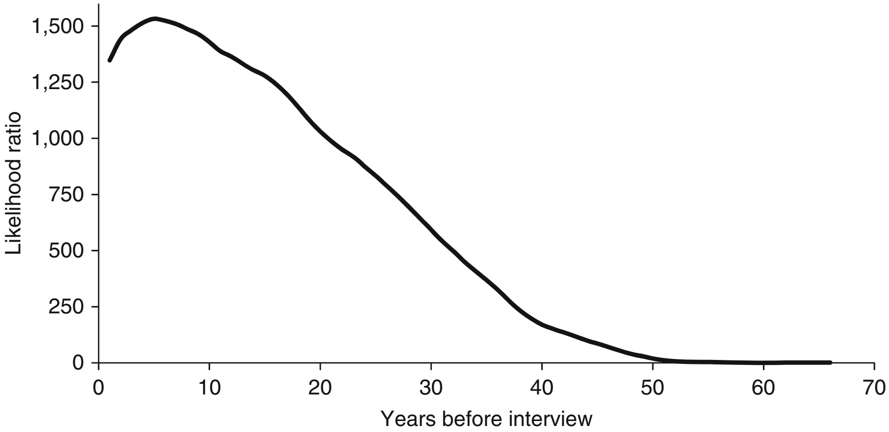


Fig. 43.10 Continuous time weighting of exposure and its impact on risk (Modified from Hauptmann et al. 2000a)

of this handbook). In general, low risks, the problem of strong confounding with other major risk factors, the problem of proper exposure assessment, and the basic assumption of the nature of the exposure-disease relationship lead to additional strategies for statistical modeling that are of special importance for environmental epidemiology.

43.3.3.1 Modeling the Exposure-Disease Function

One basic issue in the study of environmental risks is the choice of an exposure-disease function to describe the real nature of the response of an environmental hazard on human health. Two main strategies may be distinguished. If, based on former studies, on animal experiments, or on general toxicological considerations, a class of functions between exposure and disease can be specified, then a parametric version of a risk model will be suitable. This strategy may be appropriate if exposure is measured on a continuous scale, as in studies on the risks of ionizing or non-ionizing radiation or in air pollution studies. The functional type specified will be related to the preliminary considerations, but linear and log-linear parameterizations of the risk ratio are very popular in environmental studies, at least as a starting point.

The linear (excess) relative risk model may be introduced as

$$\frac{p}{1-p} = \mu \cdot (1 + \beta x), \tag{43.8}$$

where p is the risk of an interesting disease under study, μ is a multiplicative intercept, x specifies the exposure quantity, and β is the excess odds ratio, that is, the increase of the odds ratio in percent. In model (43.8), the true odds ratio equals $1 + \beta x$.

In contrast, the log-linear model may be stated as

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \gamma \cdot x, \tag{43.9}$$

where p is the risk, α is an intercept, γ is the log odds ratio, and x the exposure. Model (43.9) corresponds to the usual logistic regression model where the odds ratio is given as $\exp(\gamma \cdot x)$.

Both models (43.8) and (43.9) are very popular in environmental epidemiology for reasons of easy interpretation as well as for an easy model fit in statistical and epidemiological software packages. Likelihood-based confidence intervals as well as tests on statistical hypotheses on the parameters can be computed straightforwardly.

However, based on the range of possible exposures in the human environment, both models have to be considered very carefully. For very small exposures x, it holds that $\exp(x) \approx 1 + x$, and therefore Eq. 43.8 and Eq. 43.9 yield similar results for large parts of a general population. The risk models (43.8) and (43.9) will differ substantially for the part of a population which is exposed to a higher degree. This is demonstrated in Fig. 43.11 for the risk of residential radon and lung cancer as an example.

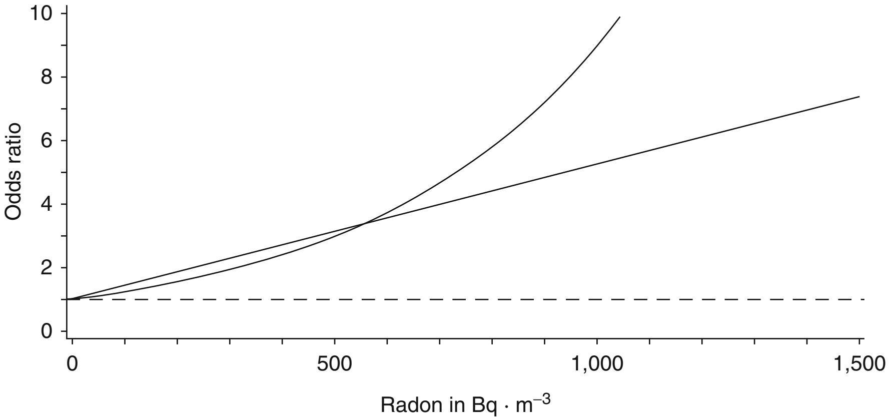


Fig. 43.11 Lung cancer risk due to residential radon; comparison of linear and log-linear risk (Modified from Oberaigner et al. 2002)

As an example, let us come back to radon exposure. It can be assumed that the exposure to radon is low for large parts of the population and that concentrations above 500 Bq·m^{-3} are unusual in the residential environment. Within this lower range, both models, the linear and the log-linear approaches, yield a similar magnitude of risk. For exposures above 500 Bq·m^{-3}, the two models will give different results that may lead to different consequences for risk management.

Therefore, non-parametric approaches are applied, especially if the real nature of the relationship is not known. On the one hand, these types of risk models may be based on categorical exposures using ordinary contingency tables. On the other hand, continuous exposure-effect relationships may be described via spline functions or fractional polynomials (cf. chapter ▶Analysis of Continuous Covariates and Dose-Effect Analysis of this handbook, Royston and Sauerbrei 2008).

There are numerous examples of this strategy, even in situations in which exposure assessment was conducted on a continuous scale and categories of exposure were defined afterward. Studies on EMF-related risks are a typical example of the use of this type of model. Figure 43.12 shows the results of a study on this topic conducted in Germany (Schüz et al. 2001).

Studies on EMF risks are usually influenced by many factors, and exposure assessment has to be conducted carefully. According to Fig. 43.12, the advantages of a model fit using categories instead of a continuous exposure assessment are obvious. While Fig. 43.12a suggests a strong increase in risk in the upper exposure category, Fig. 43.12b shows a smoother increase from one exposure category to the next. Thus, models using exposure categories increase the degrees of freedom, and any kind of exposure-disease relationship may be estimated. But this strategy is fully data-driven, and exposure-disease relationships may occur which are not plausible in any situation. For example, Ahlbom et al. (2000) reported results from a meta-analysis on EMF exposures with the same exposure categories as

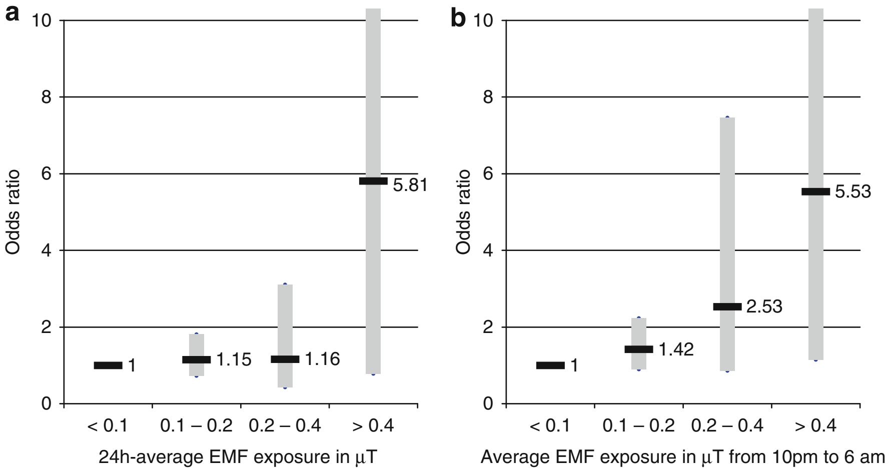


Fig. 43.12 Leukemia risk due to EMF exposure in Germany; odds ratios and 95% confidence intervals by EMF exposure in µT; (**a**) average exposure during 24 h; (**b**) average exposure during the night from 10 p.m. to 6 a.m. (Schüz et al. 2001)

in Fig. 43.12. Ahlbom et al. (2000) reported odds ratios of 1.58, 0.79, and 2.13 for the categories 0.1–0.2 μT, 0.2–0.4 μT, and more than 0.4 μT, respectively, compared to the reference category of less than 0.1 μT. These results do not agree with a monotonous exposure-disease relationship, and evidence for its true nature will be hard to obtain. However, the general strategy of using non-parametric approaches is well established in all kinds of epidemiological studies and can be considered as a basic tool for estimating risk coefficients in environmental studies.

43.3.3.2 Confounding and Interaction and Their Impact on Low Risks

Selecting the type of statistical model for investigations in environmental epidemiology is also influenced by further risk factors that may confound or modify the association under study. Given that most environmental risk factors will cause a low risk, these potentially distorting influences need careful scrutiny to avoid biases and misinterpretations of the results.

The influence of smoking in relation to the effects of environmental exposures on respiratory health may be used as a classical example. The smoking-related relative risk (RR) of lung cancer is reported to be in the order of 10 or more, in contrast to the RR of residential radon or air pollution, with RRs of less than two (Pershagen et al. 1994; Boffetta et al. 1998; Kreienbrock et al. 2001). The effect of smoking on respiratory symptoms like strong cough or obstructive bronchitis is in the range of around two to three but only in the order of 1.3 for the comparison of areas that are polluted differently (Dockery et al. 1993; Pope et al. 1995; Wolf-Ostermann et al. 1995). On the one hand, these examples suggest that it is necessary to incorporate these strong risk factors into a common risk model. On the other hand, these risk factors may dominate the model, requiring the development of proper strategies for the process of model and variable selection.

Following Greenland (1989), the selection of variables in epidemiological studies is based on two concepts: classical selection procedures such as backward or forward selection in regression models (cf. chapter ▶Regression Methods for Epidemiological Analysis of this handbook) and a sophisticated analysis of the associations and interacting mechanisms of the variables under study made in advance.

These in-advance analyses have to be conducted carefully. Sensitivity analyses can be considered as an appropriate way to deal with this problem. Here, sensitivity analyses have to be carried out of both the disease variable as well as the different risk factors under study.

Studies on respiratory health and air pollution may serve as a typical example. First, respiratory health has to be defined precisely. Often very different definitions are possible. The definition of asthma is a classical example: The US National Heart, Lung, and Blood Institute defines asthma as “a chronic inflammatory disorder of the airways [that] causes recurrent episodes of wheezing, breathlessness, chest tightness and coughing, particularly at night or in the early morning, usually associated with widespread but variable airflow obstruction that is often reversible, either spontaneously or with treatment.” In contrast the American Lung Association defines asthma as “a chronic disease of the lungs in which the airways overreact to certain factors by becoming inflamed or obstructed, making it difficult to breathe comfortably” (JHSM 2004).

This shows that varying definitions of a disease may result in a large variety of measures that will influence prevalence and incidence measures and also affect the dependent variable under study in a regression model. Therefore, it is helpful to calculate the same risk models for different definitions of the outcome to compare the patterns of the exposure-disease relationship.

If, as in most cancer studies, a symptom is defined in a clear and unequivocal way, sensitivity analyses of (concurrent) risk factors are worthwhile. Risk models can be calculated by systematically omitting one of the risk factors to assess the influence of the omitted factor by comparing the reduced model with the full model. This type of sensitivity analysis is a basic method in meta-analysis to assess the impact of each single study (cf. chapter ▶Meta-Analysis in Epidemiology of this handbook).

If a large number of risk factors and/or strong and weak risk factors are incorporated simultaneously into one model, this may yield cross-classifications with only small numbers of observations, even if the overall sample size of the study is large. This causes a reduced statistical power due to a loss of precision, as can be demonstrated by studies on the interaction of smoking and residential radon, like the Swedish nationwide study conducted by Pershagen et al. (1994) (Table 43.5).

This substantial study included 1,281 cases and 2,576 controls. Assigning these subjects to the exposure categories and especially dividing them into sub-strata related to both smoking and environmental exposure yield small sample sizes in

Table 43.5 Lung cancer risk due to smoking and residential radon in Sweden: number of cases and controls (*subj.*), odds ratios (*OR*), and 95% confidence intervals (*CI*) (Pershagen et al. 1994)

	Radon exposure in Bq m^{-3}									
	<50		50–80		80–140		140–400		>400	
Smoking status	Subj.	*OR* (CI)	Subj.	*OR* (CI)	Subj.	*OR* (CI)	Subj.	*OR* (CI)	Subj.	*OR* (CI)
	64	1	36	1.1	35	1.0	38	1.5	5	1.2
Never smoked	443	–	240	(0.7–1.7)	252	(0.6–1.5)	198	(1.0–2.3)	31	(0.4–3.1)
	35	2.6	21	2.4	24	3.2	27	4.5	1	1.1
Ex-smoker	105	(1.6–4.2)	69	(1.3–4.3)	63	(1.8–5.6)	48	(2.6–8.0)	8	(0.1–9.0)
	103	6.2	60	6.0	62	6.1	53	7.3	12	25.1
Current <10 cig. a day	128	(4.2–9.2)	79	(3.8–9.4)	79	(3.9–9.5)	59	(4.5–11.7)	4	(7.7–82.4)
	168	12.6	85	11.6	94	11.8	83	15.0	16	32.5
Current >10 cig. a day	102	(8.7–18.4)	63	(7.4–18.0)	71	(7.7–18.2)	42	(9.4–24.0)	4	(10.3–102.1)
	82	4.7	66	5.9	57	5.3	45	5.4	9	8.8
Unknown	174	(2.9–7.7)	110	(3.5–10.0)	103	(3.1–9.2)	89	(3.1–9.5)	12	(3.3–23.7)

special subclasses. For example, the number of subjects is small in the highest exposure category (more than 400 Bq·m^{-3}). Stratifying this group into subgroups by smoking status results in small numbers and leads to less precision in estimating the related effects, as is reflected by the wide confidence intervals shown in Table 43.5.

Sensitivity analyses of all possible risk factors have to be a combination of an exploratory process of exposure assessment and of the analyses of the association, confounding, and interaction mechanisms between these factors. The etiology of pediatric asthma may serve as a consolidated example of this processes as summarized by Johnson et al. (2002) (see Fig. 43.13).

The outcome of persistent (pediatric) atopic asthma has to be recognized as a complicated process involving different steps of development of the disease and common and interacting influences by different risk factors. Environmental agents may play a role as independent risk factors or as interacting and confounding variables. For example, most of the possible risk factors in the areas "residence" and "environmental hygiene" in Fig. 43.13 are more or less associated, and confounding bias may be present if this is not adequately described or modeled in an epidemiological investigation. Some studies report such associations (Weiland et al. 1994; Ponsonby et al. 2000), although the majority of studies do not support a strong relationship between air pollution and atopic asthma.

The Committee on the Assessment of Asthma and Indoor Air, Division of Health Promotion and Disease Prevention, Institute of Medicine, concluded that there is sufficient evidence for a causal relationship between exposure to allergens produced by cats, cockroaches, and house dust mites and exacerbations of asthma in sensitized individuals and between environmental tobacco smoke exposure and exacerbations of asthma in preschool-aged children. Besides these findings it was suggested that there is sufficient evidence of associations between several exposures and exacerbations of asthma, like the exposure to biological allergens produced by dogs, fungi, molds, and rhinovirus and to chemicals such as high levels of NO_2 and NO_x. Limited evidence of an association was suggested for exposures of children and adults to the biological allergens domestic birds, *Chlamydia pneumoniae, Mycoplasma pneumoniae*, or respiratory syncytial virus (RSV) or to the chemicals environmental tobacco smoke, formaldehyde, or fragrances (IOM 2000).

The evaluations of the Institute of Medicine are a synopsis of hundreds of studies on this topic. But the overall evidence of a single risk factor is influenced by a "network of associations of risk factors," from which some are identified as causal and some as associated. For example, allergens related to pets like cats and dogs may be measured and separated as factors. However, strong associations of behaviors in pet holding have to be taken into account, and it is difficult to separate them from the exposure factors above, especially if there is a large association between factors.

Therefore, powerful studies are needed both in terms of sample sizes as well as in terms of high quality in exposure assessment. Synopses like the asthma studies mentioned above, systematic reviews, meta-analyses, and pooling of studies may be useful to avoid biased conclusions about the effect of weak risk factors in environmental epidemiology.

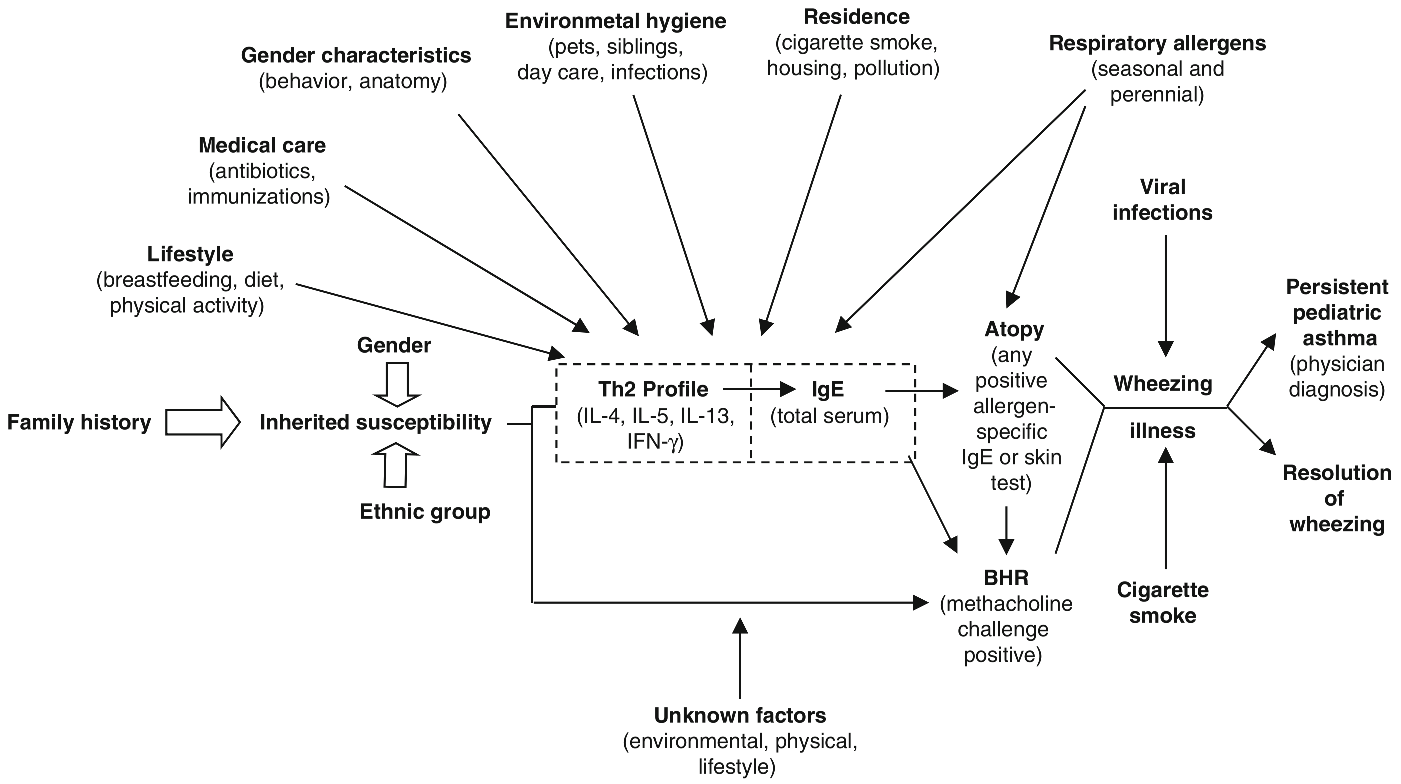


Fig. 43.13 Factors and markers potentially associated with the development of persistent pediatric atopic asthma; *TH* T-helper cell, *Ig* immunoglobulin, *IL* interleukin; *IFN-γ*s interferon gamma, *BHR* bronchial hyperactivity (Modified from Johnson et al. 2002)

43.3.3.3 Correction for Errors in Exposure Assessment

As outlined in Sect. 43.3.2, one major problem of an investigation in environmental epidemiology is the categorical or the continuous assessment of an environmental exposure. The statistical inference on this risk factor is linked to the problem of information bias if misclassification (for categorical exposures) or uncertainty (for continuous exposures) in the exposure assessment occurs.

There exist numerous examples for information biases due to misclassification in exposure assessment. Misclassification is often due to recall bias, which is typical in case-control studies. It may occur if cases report environmental exposures more often than controls, who do not care so much about environmental exposures because they are not diseased. In this situation a bias away from the null is introduced and variables may be erroneously identified as risk factors. The same direction of bias may result if interviewers tend to ask more detailed questions on exposures in cases than in controls.

Biases due to misclassification should be avoided by the design of a study and during data collection, for example, by standardization of interview techniques (cf. chapter ▶Epidemiological Field Work in Population-Based Studies of this handbook). Under certain assumptions they may be corrected by appropriate methods of adjustment for misclassification. This is usually done by estimating the sensitivity and the specificity of the categorical exposure assessment and adjusting the observed measures of disease by these estimators (cf. chapter ▶Measurement Error, see also chapter ▶Misclassification of this handbook).

But even the continuous measurement of environmental exposures may lead to uncertainty in the exposure assessment. The sources of error are numerous, like the accuracy of the detectors, the laboratory procedures, the positioning of the measurement devices, extrapolations to past exposure, or gaps in the exposure history.

These errors have to be incorporated into the statistical models used to estimate risk coefficients similar to the procedures of adjustment for misclassification. Statistical models that take uncertainty in exposure assessment into account are related to special computational efforts, and therefore different techniques were developed especially during the 1990s thanks to the overall availability of high-capacity computers (cf. also Tosteson et al. 1989; Armstrong 1990, 1998; Carroll et al. 1995; Michels 2001). Often the development of these techniques was motivated by examples from environmental epidemiology. Here, the exposure assessment is based on measurements, and the ordinary assumption of a continuous risk factor with uncertainty is fulfilled. For example, Thomas et al. (1993) applied special techniques for studies of the impact of EMF on childhood leukemia. Lagarde et al. (1997); Reeves et al. (1998), and Heid et al. (2002) investigated models for proper incorporation of the uncertainty of radon exposures and its impact on lung cancer risk. Zeger et al. (2000) and Dominici et al. (2000) examined the impact of errors in the particulate matter on mortality.

These studies show that the impact of information bias tends to increase the impact of usual random error. It is therefore of major importance to incorporate the error in the exposure assessment into the risk model. Correction for errors in

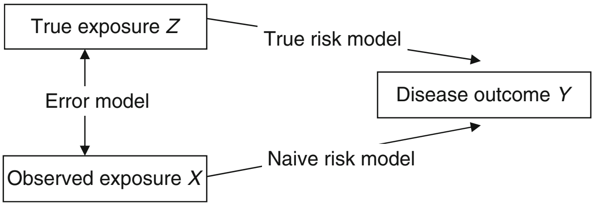


Fig. 43.14 True exposure Z and observed exposure X and their relationship to the disease outcome Y

exposure generally requires several models (cf. chapter ▶Measurement Error of this handbook): the model for the true exposure, the true exposure-disease model linking true exposure and disease, and the error model linking true exposure and measured exposure. Given these models, the exposure-disease model accounting for errors in exposure measurement can be derived by linking measured exposure and disease (Fig. 43.14).

As a general strategy, the variable X, which represents the observed result of an exposure assessment (cf. Sect. 43.3.2) is not fixed but has an error structure. The variable X has to be contrasted with the true value Z of the exposure. Following Heid (2002) in describing the deviation between X and Z, five classifying characteristics of error models may be considered: (1) random vs. systematic error, (2) non-differential vs. differential error, (3) homoscedastic vs. heteroscedastic error, (4) additive vs. multiplicative error, and (5) classical vs. Berkson error.

A random error is generally considered as a measurement error. It is characterized by an unsystematic deviation below or above the true value that average to zero. Usually all laboratory devices used in environmental measurements are prone to this error. In contrast, systematic errors lead to an overestimation or an underestimation of all individual measurements and do not average to zero. This problem may occur, if measurements were conducted on different technical levels, which is a special problem in multicenter studies or in meta-analyses. Here, intercomparison exercises of the different methods will give a deeper insight into the problem (Hollander et al. 1990; Kreienbrock et al. 1999; Wellmann et al. 2001; Bochicchio et al. 2002; Janssens 2004).

Similar to the effect of non-differential misclassification of categorical exposures, it can be assumed that error in the exposure assessment will attenuate the true exposure-disease relationship as long as this error is non-differential. A non-differential misclassification of the exposure is present if the error has the same magnitude and direction among diseased and non-diseased study participants. Otherwise (differential error), the direction of bias is not obvious in advance, and any direction is possible (cf. chapter ▶Measurement Error of this handbook). Figure 43.15 shows the effect on non-differential error for studies of lung cancer risk due to residential radon in the UK and in Sweden, where the reported excess relative risk increases by approximately 50% after adjustment for measurement error. This situation may be considered as typical in environmental epidemiology.

The problem of homoscedastic vs. heteroscedastic errors and the problem of additive or multiplicative errors are related to the nature of the statistical distributions

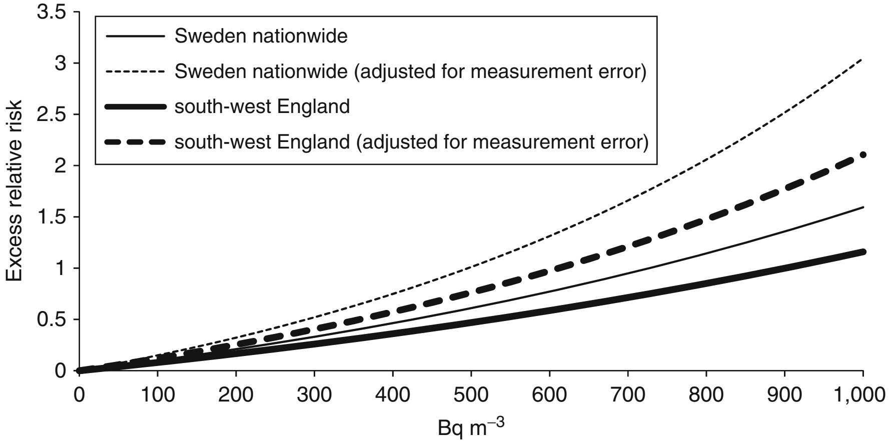


Fig. 43.15 Excess relative risk due to residential radon in the UK and in Sweden; results of classical log-linear models and adjustment for measurement error (Darby et al. 1998; Lagarde et al. 1997)

which are stated for the measurements or for the exposure in general. For an additive error, the spread of the true exposure given the measured exposure is constant for the full range of the exposure. In this situation a normal distribution is often appropriate. In contrast, for a multiplicative error, the spread increases proportional to increasing exposure. Therefore, multiplicative errors usually are assumed to be log-normal, which often is evaluated by measurement and monitoring programs on the population level. This is more or less true for many environmental agents like radon (Bäverstam and Swedjemark 1991; Gunby et al. 1993; Miles 1994; Lubin et al. 1995), classical pollutants of outdoor air (Ebelt et al. 2001; Wallace et al. 2003), endotoxin levels (Park et al. 2000), and many more.

Finally, errors of the classical type arise, if a quantity X is based on measurements by some device and repeated measurements would vary around the true value Z. This situation can be assumed for all kinds of laboratory devices, for which a measurement error is reported.

In contrast, error of the Berkson type occurs if the exposure, which is assigned to each individual, is derived from an overall group characteristic. The same approximate exposure value (proxi) X is used for all members of the group, and the true exposure Z varies randomly around this proxi with mean equal to it. This type of error may occur in a study on the effects of air pollution on respiratory health, if all study participants of a defined region are assigned to the same proxi X, for example, the result of a single measurement station. Thus, Berkson error may occur if exposure is measured via (locally) fixed monitors instead of personal detectors or if exposures have to be approximated due to missing values in the measurements (cf. chapter ▸Measurement Error of this handbook; Lagarde et al. 1997; Armstrong 1998; Reeves et al. 1998).

An overwhelming variety of models exists to estimate the effects of uncertainty of exposure assessment. Regression calibration is one of these (cf. chapter ▸Measurement Error; Rosner et al. 1989; Carroll et al. 1995). The main advantage of regression calibration is that it can be used in rather complicated measurement error models. The method includes three steps: first, find a mean (calibration) model for the true regressors Z depending on the exposure assessment X; second, fit the main model by plugging in the estimates from the calibration model. The third step is the correction of the variance estimation of the main model. Although this method was developed for multiple logistic regression accounting for errors in more than one variable, application is usually restricted to account for errors of the primary risk factor of interest. The error model, that is, the mathematical formulation of the deviation of the measured exposure X from the true exposure Z, is the most important model assumption, upon which all of the correction methods rely. In fact, assumptions about the error model are a particular source of concern regarding correction of measurement error (Michels 2001).

Therefore, substantial effort was devoted to evaluate these assumptions about the different model characteristics as well as in the different fields of applications. This was done by theoretical considerations, by simulation studies, as well as by applying different techniques on selected studies (Armstrong 1990; Thomas et al. 1993; Ibibarren et al. 1996; Lagarde et al. 1997; Reeves et al. 1998; Carrothers and Evans 2000; Dominici et al. 2000; Zeger et al. 2000; Heid et al. 2002; Field et al. 2002; Heid et al. 2006).

In conclusion, a clear distinction between the components of classical and Berkson error is essential in the assessment of error sources and for establishing an error model. This differentiation is crucial due to the different impact of these two error types. The classical error is able to induce severe bias on the risk estimate; multiplicative classical error may even distort the dose-response curve. This bias can be reduced by using the mean of multiple measurements in the analysis that require internal replicate measurements for each individual, or it can be corrected for by using the information from (internal or external) replicate measurements of a subgroup. Also the spuriously narrow confidence intervals for uncorrected risk estimates in the presence of classical error can be adjusted. Therefore, it can be recommended that more internal repeated measurements in future epidemiological studies should be conducted, for example, by using more than one detector per study participant.

At first glance, the Berkson error is less problematic in environmental epidemiology, since usually it does not induce notable bias of the risk estimates. However, Berkson error weakens the precision of the estimates and therefore leads to a loss of power that should be avoided, for example, by a proper individual exposure assessment. This will, however, introduce the classical error which may be reduced by replicate measurements, but this does not hold for the Berkson error. Simplified, classical error is related to the measurement process, whereas Berkson error is often a matter of defining the exposure groups. Using stationary monitors (e.g., using the distance of a home to the next emitter of an environmental agent instead of individual measurements) or using a person's affiliation to a group in order to use the exposure assigned to this group (e.g., using job-environment-exposure matrices instead of personal monitors) is a question of how to define the exposure group; this induces Berkson error.

The general statement that non-differential, random, and homoscedastic errors attenuate regression coefficients applies only to the classical error. To assume the sum of both error type's sizes as known and to vary the percentage of the Berkson error are one option (cf. Mallick et al. 2002). An additional option is a two-dimensional view to the measurement error, that is, a classical-type dimension and a Berkson-type dimension, where the size of each dimension needs to be studied separately. The full error is represented in the continuum of a two-dimensional space (cf. Zeger et al. 2000). Exposure assessment should therefore not only aim to be as accurate and precise as possible but also provide a model of the measurement errors that unavoidably remain even with clear differentiation of classical and Berkson components.

43.4 Conclusions

By definition, environmental epidemiology focuses on health problems due to the environment where individuals live rather than due to their personal characteristics or lifestyles. During the past centuries, there has been a remarkable discourse on environmental health and environmental epidemiology, and a huge number of individual studies as well as pooling of individual studies and meta-analyses have contributed to a large overall knowledge base on environmental hazards. As a consequence, many public health issues have been addressed by reducing contamination of air, water, soil, and food to the benefit of many parts of the world's population. One of the most prominent interventions during the last decade was the prohibition of smoking in public buildings, restaurants, and even pubs for reducing environmental tobacco smoke.

However, many multifactorial diseases are not yet fully understood, and the scientific focus has changed during the last decade, as developments in epidemiology have kept up with those in molecular biology and genetics.

A typical example is the discussion on the health impact of air pollution on respiratory diseases. One such issue is atopic asthma, and studies were conducted to find a relationship between air pollution and the incidence of the disease. But the studies failed or observed only little environmental influence. Therefore, the focus was shifted to the nature of allergic disease itself, and techniques of molecular biology and genetics have now been widely used to deepen our knowledge on this topic (cf. Johnson et al. 2002).

Therefore, the study of gene-environment interactions has been and will continue to be a major subject of epidemiological investigations, for example, for asthma, for cancer, and for other diseases to which modern molecular and genetic approaches can be applied. A serious disadvantage of all these studies is that investigations in molecular and genetic epidemiology tend to be very expensive, so that study sizes have to be restricted. For example, Kalayci et al. (2004) compared plasma levels of MCP-4 in 30 patients who presented for emergent treatment of asthma with levels in 90 subjects with chronic-stable asthma matched for age, sex, and ethnicity within an entire cohort of 596 subjects. Nowadays, however, genome-wide association studies make the necessary genetic information affordable, but the collection of environmental information is usually not part of these studies. This impedes the inclusion of genetic data in environmental studies on a broad scale.

Thus, there is a contradiction between the original idea of an environmental study with large sample sizes and extensive measurement of environmental agents on the one hand and a study in molecular and genetic epidemiology with genetic information only on the other hand. Therefore, the information content of studies on gene-environment interactions is often limited because of their insufficient ability to account for confounding.

This situation may be referred to as the "restricted information problem" in the analysis of gene-environment interactions. Two major ways may be identified for further research on this topic. To increase power, all possible measures have to be taken to improve the precision of molecular and genetic techniques and the assessment of exposures (cf. Sects. 43.3.2 and 43.3.3 and chapters ▸Molecular Epidemiology and ▸Statistical Methods in Genetic Epidemiology of this handbook). For example, first studies on the impact of environmental tobacco smoke (ETS) on respiratory diseases used urinary cotinine concentrations as a biomarker (Ehrlich et al. 1992). However, cotinine levels in the urine reflect only recent exposures and can therefore not replace exposure histories obtained via questionnaires that give a good estimate of the long-term cumulative exposure.

The second approach to cope with the "restricted information problem" is directly linked to the design and the statistical analysis of the investigation. In practice it is often impossible to detect an effect of a single agent because the various exposures are strongly correlated and the exposure that has been measured may actually act as a surrogate to the whole mixture of agents. Therefore, it is necessary to find statistical models which make use of the correlation of all possible independent risk factors, as well as all modifying and confounding variables. For example, in a study on the health impact of toxic substances ingested with food, strong correlations have to be made explicit that are due to nutritional habits, like the consumption of special types of seafood. Therefore, this has to be addressed in detail during the phase of constructing a final risk model and by appropriate specification of statistical methods (cf. chapter ▸Regression Methods for Epidemiological Analysis of this handbook).

The development of these models, besides the molecular and genetic view on a special health issue, is ordinarily linked to the personal exposures of an individual recruited for the study. On the other hand, many environmental exposures act on an aggregated level. For example, outdoor air pollution is the same, drinking water is the same, or contaminated soil often is the same for major parts of a population, for example, all inhabitants of a particular area. From the statistical point of view, this yields a Berkson-type error and a correlation of the exposures between the study participants, or even more extremely, it yields sub-classes of participants exposed in the exact same way. On the one hand, classes of hierarchical models may be used to find a proper risk model in this situation. On the other hand, the extraordinarily large number of possible environmental hazards may even force epidemiologists to conduct ecological studies to find rough estimates of risk.

In regard to the complex nature of possible environmental hazards, Pekkanen and Pearce (2001) pointed out that increasing emphasis on individual exposures, susceptibility and disease mechanisms, puts environmental epidemiologists in danger of losing their population perspective of disease causation and prevention. To avoid this and to continue the successful work of environmental epidemiology and public health, environmental health problems should be approached on four different levels: the molecular, the individual, the population, and the ecosystem level. Within and between these levels, research and development of new methods is needed, but it will be crucially important to choose the most appropriate level of research for a particular environmental problem. For example, the health impact of climate change will be one of the most important health problems of the future requiring research, but ordinary designs and exposure assessments may not be adequate to give answers to the underlying questions.

This may initiate new methodological concepts in study designs, in biological and genetic markers, in exposure assessment, as well as in the statistical analysis of studies in environmental epidemiology. In the last decades, new analytical methods in molecular biology and genetics have enriched designs and techniques in epidemiological investigations on environmental hazards. The challenge for further work is to exploit these new areas of scientific cooperation.

Summarizing, it must be determined whether there is

– Sufficient evidence of an association
– Limited or suggestive evidence of an association
– Inadequate or insufficient evidence to determine whether or not an association is present
– Limited or suggestive evidence of no association
– Evidence of no association

This judgment should be combined with an evaluation of the public health impact of an environmental problem. For this purpose, assessment of the etiological fraction due to the exposure in question may give an important input.

Acknowledgements The author likes to thank Michael Gerken, Michael Hauptmann, Iris Heid, Michaela Kreuzer, Wilhelm Oberaigner, Angelika Schaffrath Rosario, H.-Erich Wichmann, and Jürgen Wellmann for their substantial help in preparing this manuscript and Judith McAlister-Hermann for her work in reviewing. A very special gratitude for their help and patience has to be offered to the editors of this handbook, Wolfgang Ahrens and Iris Pigeot, and to the reviewers for their helpful comments.

References

Ahlbom A, Day N, Feychting M, Roman E, Skinner J, Dockerty J, Linet M, McBride M, Michaelis J, Olsen JH, Tynes T, Verkasalo PK (2000) A pooled analysis of magnetic fields and childhood leukaemia. Br J Cancer 83:692–698

Alavanja MCR, Brownson RC, Lubin JH, Berger E, Chang J, Boice JD Jr (1994) Residential radon exposure and lung cancer among nonsmoking women. J Natl Cancer Inst 86:1829–1837

Alavanja MC, Lubin JH, Mahaffey JA, Brownson RC (1999) Residential radon exposure and risk of lung cancer in Missouri. Am J Public Health 89:1042–1048

Anderson TW (1984) An introduction to multivariate statistical analysis, 2nd edn. Wiley, New York

Anderson HR, deLeon AP, Bland M, Bower JS, Strachan DP (1996) Air pollution and daily mortality in London: 1987–1992. Br Med J 312:665–669

Armstrong BG (1990) The effects of measurement errors on relative risk regression. Am J Epidemiol 132:1176–1184

Armstrong BG (1998) Effect of measurement error on epidemiological studies of environmental and occupational exposures. Occup Environ Med 55:651–656

Auvinen A, Mäkeläinen I, Hakama M, Castrén O, Pukkala E, Reisbacka H, Rytömaa T (1996) Indoor radon exposure and risk of lung cancer: a nested case-control study in Finland. J Natl Cancer Inst 88:966–972, Erratum. J Natl Cancer Inst 90:401–402 (1998)

Barros-Dios JM, Barreiro MA, Ruano-Ravina A, Figueiras A (2002) Exposure to residential radon and lung cancer in Spain: a population-based case-control study. Am J Epidemiol 156:548–555

Bäverstam U, Swedjemark G-A (1991) Where are the errors when we estimate Radon exposure in retrospect? Radiat Prot Dosim 36:107–112

Baysson H, Billon S, Laurier D, Rogel A, Tirmarche M (2003) Seasonal correction factors for estimating radon exposure in dwellings in France. Radiat Prot Dosim 104:245–252

Bell ML, Davis DL, Fletcher T (2004) A retrospective assessment of mortality from the London smog episode of 1952: the role of influenza and pollution. Environ Health Perspect 112:6–8

Blot WJ, Xu Z-Y, Boice JD, Zhao D-Z, Stone BJ, Sun J, Jing LB, Fraumeni JF (1990) Indoor radon and lung cancer in China. J Natl Cancer Inst 82:10–25

Bochicchio F, McLaughlin JP, Walsh C (2002) Comparison of radon exposure assessment results: ^{210}Po surface activity on glass objects vs contemporary air radon concentration. Radiat Meas 36:211–215

Boffetta P, Agudo A, Ahrens W, Benhamou E, Benhamou S, Darby SC, Ferro G, Fortes C, Gonzalez CA, Jöckel KH, Krauss M, Kreienbrock L, Kreuzer M, Mendes A, Merletti F, Nyberg F, Pershagen G, Pohlabeln H, Riboli E, Schmid G, Simonato L, Tredaniel J, Whitley E, Wichmann HE, Winck C, Zambon P, Saracci R (1998) Multicenter case-control study of exposure to environmental tobacco smoke and lung cancer in Europe. J Natl Cancer Inst 90:1440–1450

Brunekreef B (2003) Design of cohort studies for air pollution health effects. J Toxicol Environ Health A 66:1731–1734

Brunekreef B, Holgate ST (2002) Air pollution and health. Lancet 360:1233–1242

Carroll RJ, Ruppert D, Stefanski LA (1995) Measurement error in nonlinear models. Chapman and Hall, London

Carrothers TJ, Evans JS (2000) Assessing the impact of differential measurement error on estimates of fine particle mortality. J Air Waste Manage Assoc 50:65–74

Cohen BL (1993) Relationship between exposure to radon and various types of cancer. Health Phys 65:529–531

Cohen BL (1997) Problems in the radon vs lung cancer test of the linear no-threshold theory and a procedure for resolving them. Health Phys 72:623–628

Darby SC, Whitley E, Silcocks P, Tharkar B, Green M, Lomas P, Miles J, Reeves G, Fearn T, Doll R (1998) Risk of lung cancer associated with residential radon exposure in south-west England: a case-control study. Br J Cancer 78:394–408

Darby SC, Hill D, Doll R (2001) Radon: a likely carcinogen at all exposures. Ann Oncol 12:1341–1351

Darby S, Hill D, Auvinen A, Barros-Dios JM, Baysson H, Bochicchio F, Deo H, Falk R, Forastiere F, Hakama M, Heid I, Kreienbrock L, Kreuzer M, Lagarde F, Makelainen I, Muirhead C, Oberaigner W, Pershagen G, Ruano-Ravina A, Ruosteenoja E, Rosario AS, Tirmarche M, Tomasek L, Whitley E, Wichmann HE, Doll R (2005) Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. Br Med J 330:223–227

Dockery DW, Pope CA, Xu X, Spengler JD, Ware JH, Fay ME, Ferris BG, Speizer FE (1993) An association between air pollution and mortality in six U.S. cities. N Engl J Med 329:1753–1759

Doll R, Peto R (1976) Mortality in relation to smoking: 20 years' observations on male British doctors. BMJ 25:1525–1536

Dominici F, Zeger SL, Samet JM (2000) A measurement error model for time-series studies of air pollution and mortality. Biostatistics 1:157–175

Ebelt S, Brauer M, Cyrys J, Tuch T, Kreyling WG, Wichmann HE, Heinrich J (2001) Air quality in postunification Erfurt, East Germany: associating changes in pollutant concentrations with changes in emissions. Environ Health Perspect 109:325–333

Ehrlich R, Kattan M, Godbold J, Saltzberg DS, Grimm KT, Landrigan PJ, Lilienfeld DE (1992) Childhood asthma and passive smoking. Urinary cotinine as a biomarker of exposure. Am Rev Respir Dis 145:594–599

Ennemoser O, Ambach W, Auer T, Brunner P, Schneider P, Oberaigner W, Purtscheller F, Sting V (1994a) High indoor radon concentrations in an alpine region of western Tyrol. Health Phys 67:151–154

Ennemoser O, Ambach W, Brunner P, Schneider P, Oberaigner W, Purtscheller F, Stingl V, Keller G (1994b) Unusually high indoor radon concentrations from a giant rock slide. Sci Total Environ 151:235–240

EpiInfo^{TM} (2000) A database and statistics program for public health professionals using Windows® 95, 98, NT, and 2000 computers

Field RW, Steck DJ, Smith BJ, Brus CP, Fisher EL, Neuberger JS, Platz CE, Robinson RA, Woolson RF, Lynch CF (2000) Residential radon gas exposure and lung cancer: the Iowa Radon Lung Cancer Study. Am J Epidemiol 151:1091–1102

Field RW, Smith BJ, Steck DJ, Lynch CF (2002) Residential radon exposure and lung cancer: variation in risk estimates using alternative exposure scenarios. J Expo Anal Environ Epidemiol 12:197–203

Finkelstein MM (1991) Use of time windows to investigate lung cancer latency intervals at an Ontario steel plant. Am J Ind Med 19:229–235

Frye C, Heinrich J, Wjst M, Wichmann HE, Bitterfeld Study Group (2001) Increasing prevalence of bronchial hyperresponsiveness in three selected areas in East Germany. Eur Respir J 18:451–458

Gerken M, Kreienbrock L, Wellmann J, Kreuzer M, Wichmann HE (2000) Models for retrospective quantification of indoor radon exposure in case-control studies. Health Phys 78:268–278

Graham JD (1997) The role of epidemiology in regulatory risk assessment. Elsevier, Amsterdam

Greenland S (1989) Modeling and variable selection in epidemiologic analysis. Am J Public Health 79:340–349

Greenland S, Sheppard AR, Kaune WT, Poole C, Kelsh MA (2000) A pooled analysis of magnetic fields, wire codes, and childhood leukemia. Childhood Leukemia-EMF Study Group. Epidemiology 11:624–634

Gunby JA, Darby SC, Miles JC, Green BM, Cox DR (1993) Factors affecting indoor radon concentration in the United Kingdom. Health Phys 64:2–12

Habash RW, Brodsky LM, Leiss W, Krewski D, Repacholi M (2003a) Health risks of electromagnetic fields, part I: evaluation and assessment of electric and magnetic fields. Crit Rev Biomed Eng 31:141–195

Habash RW, Brodsky LM, Leiss W, Krewski D, Repacholi M (2003b) Health risks of electromagnetic fields, part II: evaluation and assessment of radio frequency radiation. Crit Rev Biomed Eng 31:197–254

Hansen MH, Hurwitz WN (1946) The problem of non-response in sample surveys. JASA 41:517–529

Hauptmann M, Lubin JH, Rosenberg PS, Wellmann J, Kreienbrock L (2000a) The use of sliding time windows for the exploratory analysis of temporal effects of smoking histories on lung cancer risk. Stat Med 19:2185–2194

Hauptmann M, Wellmann J, Lubin JH, Rosenberg PS, Kreienbrock L (2000b) Analysis of exposure-time-response relationships using a spline weight function. Biometrics 56:1105–1108

Heid IM (2002) Measurement error in exposure assessment: an error model and its impact on studies on lung cancer and residential radon exposure in Germany. Doctoral Thesis Ludwig-Maximilians-Universität München

Heid IM, Küchenhoff H, Wellmann J, Gerken M, Kreienbrock L, Wichmann HE (2002) On the potential of measurement error to induce differential bias on odds ratio estimates: an example from radon epidemiology. Stat Med 21:3261–3278

Heid IM, Schaffrath Rosario A, Kreienbrock L, Küchenhoff H, Wichmann HE (2006) The impact of measurement error on studies on lung cancer and residential radon exposure in Germany. J Toxicol Environ Health A 69:701–721

Heidrich J, Wellmann J, Heuschmann PU, Kraywinkel K, Keil U (2007) Mortality and morbidity from coronary heart disease attributable to passive smoking. Eur Heart J 28:2498–2502

Heinrich J, Hoelscher B, Wichmann HE (2000) Decline of ambient air pollution and respiratory symptoms in children. Am J Respir Crit Care Med 161:1930–1936

Heinrich J, Hoelscher B, Frye C, Meyer I, Wjst M, Wichmann HE (2002) Trends in prevalence of atopic diseases and allergic sensitization in children in Eastern Germany. Eur Respir J 19:1040–1046

Heuschmann PU, Heidrich J, Wellmann J, Kraywinkel K, Keil U (2007) Stroke mortality and morbidity attributable to passive smoking in Germany. Eur J Cardiovasc Prev Rehabil 14: 793–795

Hoek G, Brunekreef B, Verhoeff A, van Wijnen J, Fischer P (2000) Daily mortality and air pollution in The Netherlands. J Air Waste Manage Assoc 50:1380–1389

Hollander W, Morawietz G, Bake D, Laskus L, van Elzakker BG, van der Meulen A, Zierock KH (1990) A field intercomparison and fundamental characterization of various dust samplers with a reference sampler. J Air Waste Manage Assoc 40:881–886

IARC, International Agency on Research on Cancer (2001) Ionizing radiation, part 2: Some internally deposited radionuclides. IARC monographs on the evaluation of carcinogenic risks to humans, vol 78. International Agency on Research on Cancer, Lyon

Ibibarren C, Sharp D, Burchfield CM, Ping S, Dwyer JH (1996) Association of serum total cholesterol with coronary disease and all-cause mortality: multivariate correction for bias due to measurement error. Am J Epidemiol 143:463–471

ICRP, International Commission on Radiological Protection (1993) Protection against radon 222 at home and at work. ICRP Publication 65. Annals of the ICRP, vol. 23, No 2. Didcot, Oxon

IOM, Institute of Medicine (2000) Clearing the air: asthma and indoor air exposures. National Academic Press, Washington DC

Janssens A (2004) Environmental radiation protection: philosophy, monitoring and standards. J Environ Radioact 72:65–73

JHSM (2004) Division of Pulmonary and Critical Care Medicine, Johns Hopkins School of Medicine. http://www.hopkins-lungs.org/programs/asthma/. Accessed 8 May 2004

Johnson CC, Ownby DR, Zoratti EM, Hensley Alford S, Williams LK, Joseph CLM (2002) Environmental epidemiology of pediatric asthma and allergy. Epidemiol Rev 24:154–175

Kalayci O, Sonna LA, Woodruff PG, Camargo CA Jr, Luster AD, Lilly CM (2004) Monocyte chemotactic protein-4 (MCP-4; CCL-13): a biomarker of asthma. J Asthma 41:27–33

Katsouyanni K, Touloumi G, Spix C, Schwartz J, Balducci F, Medina S, Rossi G, Wojtyniak B, Sunyer J, Bacharova L, Schouten JP, Ponka A, Anderson HR (1997) Short-term effects of ambient sulphur dioxide and particulate matter on mortality in 12 European cities: results from time series data from the APHEA project. Air pollution and health: A European approach. Br Med J 314:1658–1663

Kleinbaum DG, Kupper LL, Morgenstern H (1982) Epidemiologic research. Van Nostrand Reinhold, New York

Kreienbrock L, Siehl A (1996) Multiple statistische Analyse von Radon-Erhebungsmessungen in der Bundesrepublik Deutschland. In: Siehl A (ed) Umweltradiaoaktivität – Geologie und Ökologie im Kontext. Ernst & Sohn, VCH, Berlin, pp 299–310

Kreienbrock L, Poffijn A, Tirmarche M, Feider M, Kies A, Darby SC (1999) Intercomparison of passive Rn-detectors under field conditions in epidemiological studies. Health Phys 76:558–563

Kreienbrock L, Kreuzer M, Gerken M, Dingerkus G, Wellmann J, Keller G, Wichmann HE (2001) Case-control study on lung cancer and residential radon in West Germany. Am J Epidemiol 153:42–52

Kreienbrock L, Pigeot I, Ahrens W (2012) Epidemiologische Methoden, 5th edn. (in German). Spektrum, Heidelberg

Kreuzer M, Heinrich J, Wölke G, Schaffrath Rosario A, Gerken M, Wellmann J, Keller G, Kreienbrock L, Wichmann HE (2003) Residential radon and risk of lung cancer in Eastern Germany. Epidemiology 14:1–10

Krewski D, Burnett RT, Goldberg MS, Hoover K, Siemiatycki J, Abrahamowicz M, White WH (2004) Validation of the Harvard Six Cities Study of particulate air pollution and mortality. N Engl J Med 350:198–199

Krewski D, Lubin JH, Zielinski JM, Alavanja M, Catalan VS, Field RW, Klotz JB, Létourneau EG, Lynch CF, Lyon JL, Sandler DP, Schoenberg JB, Steck DJ, Stolwijk JA, Weinberg C, Wilcox HB (2006) A combined analysis of North American case-control studies of residential radon and lung cancer. J Toxicol Environ Health A 69:533–597

Lagarde F, Pershagen G, Akerblom G, Axelson O, Bäverstam U, Damber L, Enflo A, Svartengren M, Swedjemark GA (1997) Residential radon and lung cancer in Sweden: risk analysis accounting for random error in the exposure assessment. Health Phys 72:269–276

Lagarde F, Axelsson G, Damber L, Mellander H, Nyberg F, Pershagen G (2001) Residential radon and lung cancer among never-smokers in Sweden. Epidemiology 12:396–404

Létourneau EG, Krewski D, Choi NW, Goddard MJ, McGregor RG, Zielinski JM, Du J (1994) Case-control study of residential radon and lung cancer in Winnipeg, Manitoba, Canada. Am J Epidemiol 140:310–322

Locher W, Unschuld PU (1999) Geschichtliches zur Umweltmedizin. In: Wichmann HE, Schlipköter HW, Fülgraff G (eds) Handbuch der Umweltmedizin. ecomed, Landsberg/Lech, pp II-1.1–II-1.12

Lubin JH (1988) Models for the analysis of radon-exposed populations. Yale J Biol Med 61:195–214

Lubin JH, Steindorf K (1995) Cigarette use and the estimation of lung cancer attributable to radon in the United States. Radiat Res 141:79–85

Lubin JH, Samet JM, Weinberg C (1990) Design issues in epidemiologic studies of indoor exposure to Rn and risk of lung cancer. Health Phys 59:807–817

Lubin JH, Boice JD, Edling CH, Hornung R, Howe G, Kunz E, Kusiak A, Morrison HI, Radford EP, Samet JM, Tirmarche M, Woodward A, Xiang YS, Pierce DA (1994) Radon and lung cancer risk: a joint analysis of 11 underground miners studies. US National Institutes of Health. NIH publication No 94–3644

Lubin JH, Boice JD Jr, Samet JM (1995) Errors in exposure assessment, statistical power and the interpretation of residential radon studies. Radiat Res 44:329–341

Mallick B, Hoffmann FO, Carroll RJ (2002) Semiparametric regression modeling with mixtures of Berkson and classical error, with application to fallout from the Nevada test site. Biometrics 58:13–20

Menzler S, Piller G, Gruson M, Rosario AS, Wichmann HE, Kreienbrock L (2008) Population attributable fraction for lung cancer due to residential radon in Switzerland and Germany. Health Phys 95:179–189

Michels KB (2001) A renaissance for measurement error. Int J Epidemiol 30:421–422

Miles JCH (1994) Mapping the proportion of the housing stock exceeding a radon reference level. Radiat Prot Dosim 56:207–210

NRC, National Research Council (1991) Environmental epidemiology: public health and hazardous wastes. National Academy Press, Washington DC

NRC, National Research Council (1999) Health effects of exposure to radon, BEIR VI. Committee on health risks of exposure to radon. Board on Radiation Effects Research, Commission on Life Science. National Academy Press, Washington DC

Oberaigner W, Kreienbrock L, Schaffrath Rosario A, Kreuzer M, Wellmann J, Keller G, Gerken M, Langer B, Wichmann HE (2002) Radon und Lungenkrebs im Bezirk Imst/Österreich. Fortschritte in der Umweltmedizin. Ecomed Verlagsgesellschaft, Landsberg am Lech

Park JH, Spiegelman DL, Burge HA, Gold DR, Chew GL, Milton DK (2000) Longitudinal study of dust and airborne endotoxin in the home. Environ Health Perspect 108:1023–1028

Pekkanen J, Pearce N (2001) Environmental epidemiology: challenges and opportunities. Environ Health Perspect 109:1–5

Pershagen G, Liang ZH, Hrubec Z, Svensson C, Boice JD (1992) Residential radon exposure and lung cancer in women. Health Phys 63:179–186

Pershagen G, Akerblom G, Axelson O, Clavensjö B, Damber L, Desai G, Enflo A, Lagarde F, Mellander H, Svartengren M, Swedjemark GA (1994) Residential radon exposure and lung cancer in Sweden. N Engl J Med 330:159–164

Pinel J, Fearn T, Darby SC, Miles JCH (1994) Seasonal correction factors for indoor radon measurements in the United Kingdom. Radiat Prot Dosim 58:127–132

Ponsonby AL, Couper D, Dwyer T, Carmichael A, Kemp A, Cochrane J (2000) The relation between infant indoor environment and subsequent asthma. Epidemiology 11:128–135

Pope CA 3rd, Thun MJ, Namboodiri MM, Dockery DW, Evans JS, Speizer FE, Heath CW Jr (1995) Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. Am J Respir Crit Care Med 151:669–674

Prüss-Üstün A, Mathers C, Corvolán C, Woodward A (2003) Introduction and methods: assessing the environmental burden of disease at national and local levels. WHO Environmental Burden of Disease Series, No. 1. World Health Organization, Geneva

Reeves GK, Cox DR, Darby SC, Whitley E (1998) Some aspects of measurement error in explanatory variables for continuous and binary regression models. Stat Med 17:2157–2177

Roemer WH, van Wijnen JH (2001) Daily mortality and air pollution along busy streets in Amsterdam, 1987–1998. Epidemiology 12:649–653

Rosner B, Willett WC, Spiegelman D (1989) Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med 8:1051–1069

Royston P, Sauerbrei W (2008) Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Wiley, Chichester

Ruosteenoja E, Mäkeläinen I, Rytömaa T, Hakulinen T, Hakama M (1996) Radon and lung cancer in Finland. Health Phys 71:185–189

Samet JM, Stolwijk J, Rose S (1991) International workshop on residential radon-epidemiology. Health Phys 60:223–227

Samet JM, Zeger SL, Dominici F, Curriero F, Coursac I, Dockery DW, Schwartz J, Zanobetti A (2000) The national morbidity, mortality, and air pollution study, part II: morbidity and mortality from air pollution in the United States. Res Rep Health Eff Inst 94:5–70; discussion 71–79

Schoenberg JB, Klotz JB, Wilcox HB, Nicholls GP, Gil-del-Real MT, Stemhagen A, Mason TJ (1990) Case-control study of residential radon and lung cancer among New Jersey women. Cancer Res 50:6250–6254

Schüttmann W (1992) Das Radonproblem im Bergbau und in Wohnungen – Historische Aspekte. In: Reiners C, Streffer C, Messerschmidt O (eds) Strahlenrisiko durch Radon. Gustac Fischer, Stuttgart/Jena/New York, pp 5–24

Schüz J, Grigat JP, Brinkmann K, Michaelis J (2001) Residential magnetic fields as a risk factor for childhood acute leukemia: results from a German population based case-control study. Int J Cancer 91:728–735

Slimani N, Kaaks R, Ferrari P, Casagrande C, Clavel-Chapelon F, Lotze G, Kroke A, Trichopoulos D, Trichopoulou A, Lauria C, Bellegotti M, Ocke MC, Peeters PH, Engeset D, Lund E, Agudo A, Larranaga N, Mattisson I, Andren C, Johansson I, Davey G, Welch AA, Overvad K, Tjonneland A, Van Staveren WA, Saracci R, Riboli E (2002) European Prospective Investigation into Cancer and Nutrition (EPIC) calibration study: rationale, design and population characteristics. Public Health Nutr 5:1125–1145

Steindorf K, Lubin JH, Wichmann HE, Becher H (1995) Lung cancer deaths attributable to indoor radon exposure in West Germany. Int J Epidemiol 24:485–492

Stidley AC, Samet JM (1993) A review of ecologic studies of lung cancer and indoor radon. Health Phys 65:234–251

Thomas D, Stram D, Dwyer J (1993) Exposure measurement error: Influence on exposure-disease relationships and methods of correction. Annu Rev Public Health 14:69–93

Tomášek L, Müller T, Kunz E, Heribanová A, Matzner J, Plaček V, Burian I, Holeček J (2001) Study of lung cancer and residential radon in the Czech Republic. Centr Eur J Public Health 9:150–153

Tosteson TD, Stefanski LA, Schafer DW (1989) A measurement-error model for binary and ordinal regression. Stat Med 8:1139–1147

UNSCEAR, United Nations Scientific Committee on the Effects of Atomic Radiation (2000) Sources and effects of ionizing radiation. UNSCEAR 2000 report to the general assembly, with scientific annexes. Vol. I: Sources. United Nations, New York

Vedal S, Brauer M, White R, Petkau J (2003) Air pollution and daily mortality in a city with low levels of pollution. Environ Health Perspect 111:45–51

Wallace LA, Mitchell H, O'Connor GT, Neas L, Lippmann M, Kattan M, Koenig J, Stout JW, Vaughn BJ, Wallace D, Walter M, Adams K, Liu LJS (2003) Particle concentrations in inner-city homes of children with Asthma: the effect of smoking, cooking, and outdoor pollution. Environ Health Perspect 111:1265–1272

Wang Z, Lubin JH, Wang L, Zhang S, Boice JD Jr, Cui H, Zhang S, Conrath S, Xia Y, Shang B, Brenner A, Lei S, Metayer C, Cao J, Chen KW, Lei S, Kleinerman RA (2002) Residential radon and lung cancer risk in a high-exposure area of Gansu Province, China. Am J Epidemiol 155:554–564

Warner KE, Mendez D, Courant PN (1996) Toward a more realistic appraisal of the lung cancer risk from radon: The effects of residential mobility. Am J Public Health 86:1222–1227

Weiland SK, Mundt KA, Ruckmann A, Keil U (1994) Self-reported wheezing and allergic rhinitis in children and traffic density on street of residence. Ann Epidemiol 4:243–247

Wertheimer N, Leeper E (1979) Electric wiring configurations and childhood cancer. Am J Epidemiol 109:273–284

Wichmann HE, Kreienbrock L, Kreuzer M, Gerken M, Dingerkus G, Wellmann J, Keller G (1998) Lungenkrebsrisiko durch Radon in der Bundesrepublik Deutschland (West). ecomed, Landsberg/Lech

Webster's Encyclopedic Unabridged Dictionary of the English Language (1989) Portland House, New York

Wellmann J, Miles J, Kreienbrock L (2001) Identification of outliers in an international radon intercomparison exercise. In: Kunert J, Trenkler G (eds) Mathematical statistics with applications in biometry. Festschrift in Honour of Prof. Dr. Siegfried Schach. Josel Eul, Lohmar/ Köln, pp 253–262

WHO, World Health Organization (2000) Air quality guidelines for Europe, 2nd edn. WHO Regional Publications, European Series No 91. WHO, Regional Office for Europe, Copenhagen

Wolf-Ostermann K, Luttmann H, Treiber-Klötzer C, Kreienbrock L, Wichmann HE (1995) Kohortenstudie zu Atemwegserkrankungen und Lungenfunktion bei Schulkindern in Südwestdeutschland – Teil 3: Einfluß von Rauchen und Passivrauchen. Zentralblatt für Hygiene und Umweltmedizin 197:459–488

Zeger SL, Thomas D, Dominici F, Samet JM, Schwartz J, Dockery D, Cohen A (2000) Exposure measurement error in time-series studies of air pollution: concepts and consequences. Environ Health Perspect 108:419–426

Nutritional Epidemiology 44

Heiner Boeing and Barrie M. Margetts

Contents

H. Boeing (✉)
Department of Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany

B.M. Margetts
Faculty of Medicine, Public Health Nutrition, University of Southampton, Southampton, UK

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_26,

44.1 Introduction

> The main objective of nutritional epidemiological research is to provide the best possible scientific evidence to support an understanding of the role of nutrition in the causes and prevention of ill-health
>
> Margetts and Nelson (1997)

Basic textbooks describing the field of nutritional epidemiology were available in the 1990s (Margetts and Nelson 1997; Willett 1998), and one of these (Willett 2013) has recently been updated. It is not the intention of this chapter to repeat all that is available in these textbooks but to highlight some key concepts and recent developments relevant to readers of a general epidemiology textbook.

Nutritional epidemiology primarily seeks to describe the distribution and variation in the nutritional behavior of individuals and relates that behavior to health outcomes. Nutritional epidemiological approaches are appropriate for a number of different purposes. Primarily, nutritional epidemiological studies seek to understand the role that nutritional factors play in causing various health outcomes. In generic terms, the aim is to explore exposure-outcome relationships. Outcomes can be defined in many different ways but include the occurrence of measures of well-being, or a disease, or intermediate disease markers, or measures of metabolic disturbance, or even measures of nutritional status such as body composition (body mass index used to define obesity, and stunting and wasting related to age-specific weight and height). Nutritional exposure is a generic term used here to describe different aspects of dietary and nutritional behavior and can include dietary intakes, dietary habits, knowledge and attitudes about food, biochemical markers, body composition, or even clinical signs of nutritional problems.

The principles of nutritional epidemiology can also be applied to studies that seek to monitor the population for the purposes of identifying individuals at risk of deficiency or excess or for describing the proportions of the population at risk. Such studies provide a basis for deciding public health actions and for informing progress on policies aimed at reducing nutrition-related problems.

Nutritional epidemiological studies follow the general principles of all epidemiological studies to develop a clear and testable research question and to provide an unbiased answer to that question. These general principles are covered elsewhere in Parts I (Concepts and Designs in Epidemiology) and II (Methodological Approaches in Epidemiology) of the handbook. The issue with assessing outcomes is not specific to nutritional studies and will not be discussed in this chapter.

The aim of this chapter is to focus on the challenges and best practice around assessing nutritional exposures. The best approach depends on the purpose of the study, and methods used to derive measures need to be fit for purpose. For example, if an individual level of accuracy is required, where within-person variation can be characterized, a simple questionnaire that assesses usual intake will not be adequate. On the other hand, if all that is required is an estimate of the population average, a less precise measure may suffice, provided the sample size and sampling frame are appropriate. Whenever nutritional data are collected, a measure of the relative validity of that measure for the purpose intended should be provided. As in all epidemiological studies, the key challenge is to avoid bias, both in terms of information and sampling, while managing measurement error that is unavoidable because of the nature of variation in behavior and the imprecise methods available.

An evidence-based approach to the development of future research priorities and public health actions depends on that evidence being available and being of sufficient quality to be useful. Similarly monitoring the effectiveness of interventions or actions aimed at improving nutrition depends on the appropriate use and availability of reliable measures in the population and among vulnerable groups. Methods that provide measures may not be suitable for all sectors of a population, and this needs to be taken into account when developing and using nutritional epidemiological tools. It is the aim of this chapter to shed some light on how best to provide this evidence.

This chapter describes the measurement of dietary exposure, the definition of nutritional exposures, measurement error and calibration of dietary intake, the statistical approaches to obtain proper estimates of dietary intake data, and relative risk estimates including energy adjustment and dietary pattern analysis. It also illustrates the organization and presentation of dietary data including implications for meta-analysis and reviews and depicts the challenge of nutrition surveys and nutritional epidemiology in public health practice.

General Considerations in Measuring Exposure The most commonly used measure of nutritional exposure is the assessment of diet. The measure may be dietary patterns, intake of a particular food, the nutrient content of a food or the whole diet, or consumption of supplements or other added substances such as additives or preservatives. Where nutrient intake is required, the relevant food composition data must be available.

The level of detail required to assess dietary exposure depends on the exposure of interest, and this in turn determines the level of accuracy required to derive the relevant measure of exposure. In selecting the approach to measuring dietary exposure, it is important to consider the relevant time frame for the assessment; is it usual long-term intake, or total life time exposure, current exposure or immediate past exposure, or exposure at a critical point in time based on some biological insight about causation? It is also necessary to consider how the measure will be expressed, as the average exposure for the individual or population or proportion above a threshold or cut-off for either the individual or the population.

Before deciding on a method of assessment, it is thus critical to have a good idea of the question being asked and the level of detail required to correctly characterize the relevant exposure. The within-person variation in exposure behavior, the between-person variation, the measurement error of the instrument, and method itself all need to be appropriately taken into account.

It is essential to develop and use a written protocol to ensure that methods are used appropriately. Observers should be trained prior to the study beginning. Part of this training should ensure that within- and between-observer error is within acceptable limits (maximum 10%) and that there is no bias in the way observers use the instruments. If there is unmeasured between-observer variation, it will be impossible to disentangle this from center-to-center variation. Advanced computer programs that guide the interview such as the multiple-pass method developed in the USA (Convey et al. 2004) or EPIC-Soft developed in Europe (Slimani et al. 2000) have helped to minimize between-observer effects.

Practical considerations of time and money may limit the choice of methods used and the impact that this may have on the results obtained needs to be taken into account. Small, but important differences may be missed by imprecise methods.

44.2 Measuring Dietary Exposure

Individual Level of Assessment When the exposure of interest is at the level of the individual, the key question is whether the measure required is for usual intake or current intake (Illner et al. 2010). Usual intake refers to average long-term habitual intake, often expressed as diet over a year. From the statistical point of view, it is sufficient to capture a limited number of days from that period to represent usual intake (see Sect. 44.4). These days should be selected at random (Fig. 44.1). The length of time required to assess both usual and current intake depends on the variety of foods consumed by the population group of interest and a consideration of determinants of dietary intake that may have changed over time.

The more the diet and the more the eating behavior vary over time, the longer the time frame that is required to capture these variations. Where it is known that environmental factors have influenced access to food at a particular time this needs to be taken into account if exposure at that time is considered to be critical. This might be the case when food supply was impaired, e.g., by drought or war. Urbanization may also affect dietary behavior. Moving from rural areas where people mostly rely on foods they have grown themselves, to urban areas where they will mostly be buying foods will have a major impact on dietary behavior. Where foods are prepared and consumed away from the home, this may also alter the composition and quality of the diet. Other important considerations may be personal characteristics of the subjects being investigated. An important aspect on which information should be obtained is the literacy of the subjects to ensure that the chosen method is suitable for the target population. In studies with deprived populations, for example, interviewer-based methods could be selected instead of self-administered questionnaires.

For current intake, the key challenge is to decide how many days of recording are required to capture the true within-person variability. Where food patterns are changing rapidly, and where the composition of the foods being eaten is changing, the methods used to assess behavior must be sensitive to these dynamics.

Population Level of Assessment Where the primary requirement of the measure of exposure is to assess population or subgroup average exposure, the main consideration will be having a suitable large and representative sample to characterize the variation of behavior in the population. The recorded exposure for any one individual will not accurately reflect their true long-term exposure, but the assumption is made that on average those overreporting exposure will be balanced out by those underreporting their exposure. This assumes random error only; if

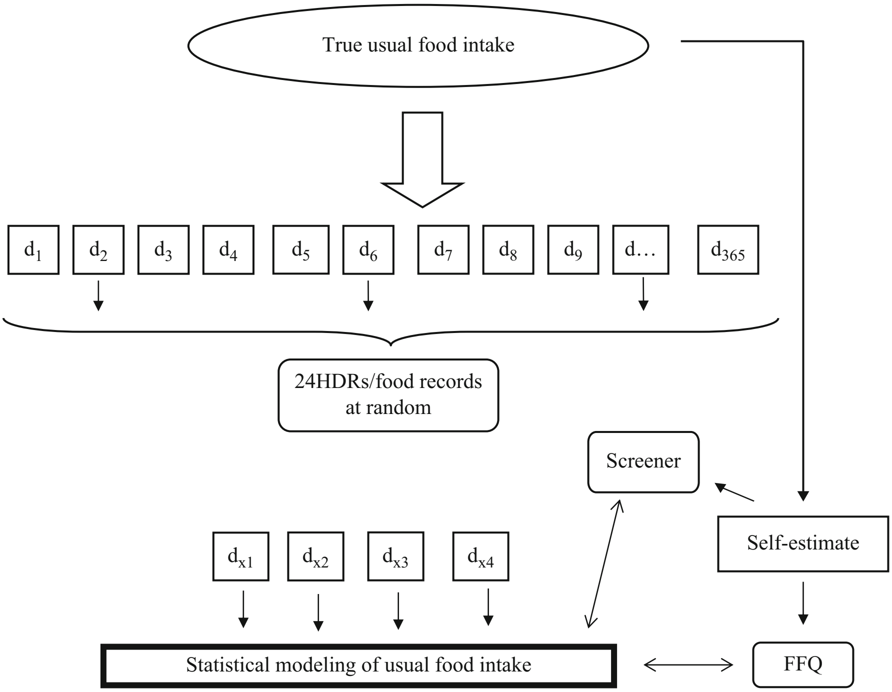


Fig. 44.1 Approaches of dietary assessment. Food and nutrient intake is realized by intake per day over life (24 HDR = 24 hour dietary recall). Usually, in studies, a year is of interest. This can be considered as true usual intake. We can estimate this intake by selecting a number of days at random in which intake is recorded. Another approach is to estimate usual intake directly via self-report. The food frequency questionnaire (FFQ) is such an instrument. Screeners are tools that address only a specific aspect of intake but not the total diet. They may be designed to assess day-to-day behavior but also to self-report usual diet. Statistical tools are now available to model usual food intake from day-to-day information. Some of them are also utilizing self-estimates in addition to day-to-day information to estimate usual food intake

individuals with certain characteristics systematically under- or overreport their true behavior, then the measured population average exposure will be biased. One such factor that may distort the measure of exposure is a person's body mass index. How to explore this will be covered further in a later section.

There is a large body of statistical data on national food availability and food availability on retail and household levels (Table 44.1). These are aggregate data to characterize regions which may serve ecological comparisons. The national statistics on food availability are systematically collected from the national statistical offices in a standardized way and prepared by the FAO in Rome. They provide an excellent overview of the global situation of food availability in the regions of the world over time (FAO 2013). Similarly, data on purchases at the household level have been used to describe the level of food security. In the European countries, household data have also been used to describe dietary practices (Naska et al. 2007).

Table 44.1 Methods of dietary assessment by type of epidemiological study (Modified from Margetts et al. 2003)

Epidemiological study	Level of aggregation required; expression of information (comment)	Method (comment)
Population or household level		
Aggregate population/ecological	• Average per capita intake compared across countries or regions or households • Trends over time within a country or region or household	Food disappearance data Food balance sheets from statistical offices Household budget surveys
Community experiment or intervention	Group level of analysis: compare outcomes for different exposures	Household budget surveys or sentinel assessment of representative individuals
Individual level		
Cross-sectional (prevalence survey)	Absolute amount	Food records or recalls (multiple days to describe within-person variation; number depends on measure required) Biochemical markers in blood and urine (check dose-response relationship with diet)
	Ranking	Food records (usually single days) 24-h recall (single or multiple days) FFQ, screeners Nutritional biochemical markers
Case-control	Past exposure at time of initiation or as proxy for past; usually categorized but may use continuous measure. (Absolute accuracy not required; assess misclassification and take into account)	FFQ of present or past diet Diet history Screeners
Cohort	Subjects intake at start of study ranked and categorized (expressed as exposed, not exposed) but may use continuous measure (absolute accuracy not required; assess misclassification and take into account, unless absolute risk to be described)	Food records 24-h recall FFQ Screeners Nutritional biochemical markers in blood or urine (does it relate to diet in sensitive manner?)
Experimental study	Individual level (usually needs to be accurate at absolute level)	Food records (multiple days to assess within-person variation) 24-h recalls FFQ (for patterns or group level comparison) Diet quality indices Nutritional biochemical markers

FFQ food frequency questionnaire

Impact of Study Design There is also a relation between type of study design and the dietary assessment instruments to be used. The major difference is between prospective and retrospective study designs. Retrospective study designs such as the case-control study need to apply retrospective dietary assessment methods since diet before onset of the disease is to be measured. Corresponding estimates of nutritional exposure are quite limited with regard to validity and accuracy of quantitative dietary intake measures.

In the next paragraphs we will summarize the key issues in assessing diet using different methods. Since there is an increasing interest in novel technologies and the use of web-based instruments (Illner et al. 2012), each section contains a description of the traditional approach and novel techniques. New information technologies will enable nutritional epidemiologists to rethink the way they currently measure diet, and this may lead to better quality information. However, most methods still require the active participation of the study subject and nutritional exposure assessment has to rely on self-reported information rather than on objectively measured parameters.

Dietary assessment methods with emphasis on novel approaches have recently been reviewed (Ngo et al. 2009; Illner et al. 2012). Particularly helpful is also the website of the risk factor monitoring and methods branch of the National Cancer Institute (NCI) as part of the National Institutes of Health of the USA (National Cancer Institute 2013) and the website of the Medical Research Council (MRC) of the UK (Medical Research Council 2013).

44.2.1 Food Record

The food record usually seeks to register prospectively all the food, drink, and supplement intake consumed over a period of time, usually broken down to 24-h periods. For some studies, where there are very specific foods of interest, the recording may be restricted to these foods or food groups. The required number of days of recording depends on how the data are to be used. For population estimates, two recorded days per person will suffice as long as the sampling covers all days of the week and all weeks of the year. Where an accurate estimate of individual exposure is required more recorded days will be needed depending on the sources of variation in the population and the exposure of interest (see Sect. 44.3). For example, if the exposure of interest is a nutrient found in a limited number of foods that are rarely eaten, then many days of recording in a large sample are required.

In theory, a major advantage of food records is that the open-ended nature of the collection allows many different exposure measures to be derived. When new exposures of interest arise, these can be derived from the records provided that such details were recorded.

Ideally consumption is recorded as it occurs, not recalled at the end of the day. It is generally the aim to record all foods eaten with a detailed description including quantity and method of preparation. Greater accuracy is usually obtained when the foods are weighed before they are consumed (weighed food record), rather than this amount being estimated or weighed before preparation.

A food record, if properly conducted, is often considered as the reference method against which less accurate methods are compared (see Sect. 44.2.4). Caution is required when making the assumption that a food record represents true intake. People often change their diets when asked to record their intake, either because the method is too time-consuming and people thus eat simpler foods or because they preferably choose foods that are socially desirable. The method relies on the participant being honest and retaining his/her usual food choices. Otherwise, it may be that technically accurate measures are obtained which are, however, biased in comparison to what a researcher would observe with an unobtrusive method.

44.2.1.1 Traditional Food Record

The food record is an open-ended method, either self-completed or administered by an interviewer. Traditionally the method requires the individual to write down what they have consumed over 24 h, either on a blank sheet or sometimes structured around meal-times or times of the day, depending on what is most appropriate for the target population. The interview approach may be conducted face-to-face or over the telephone. The researcher does not have to make assumptions as to which foods to include. This is also a disadvantage since correct assignment and coding of the foods are a major time-consuming effort, in particular if the subject has not given the respective details and/or has consumed many composite dishes. Technically, a food record can be collected, stored, and analyzed later (Bingham et al. 2003).

As an alternative to writing down what was eaten, it is possible to ask subjects to use a voice recorder; the recoding has then to be transcribed and to be processed further (Kohlmeier 1995; Illner et al. 2012).

If the study aims to record group mean intake and assuming the errors in the study sample are randomly distributed, a 1-day record will provide a reasonable estimate of the group mean. Where the requirement is to measure an individual's dietary intake and/or distributions, then more days of recording per subject will be required, depending on the extent of the within-person variability of the nutritional exposure of interest. Micronutrients that are found in large amounts in foods rarely eaten (vitamin A in organ meats) will require more days (perhaps up to 30 days) of recording to capture those days when the nutrient-rich foods are eaten. Macronutrients are generally supplied from a wide range of foods, and an individual's usual intake can probably be derived from 3 to 4 days of recording. In practice, researchers tend to ask participants to record more than 1 day, such as 4- or 7-day records, at one occasion. Often, it seems useful to discard the first day due to the fact that this day often differs from the remaining days. Also in terms of the statistical properties, a day from a consecutive record is not a random selection of possible days because a day out of a series of days has a high chance of being correlated with the other days, including the error (one day with a large intake, the other day with a small intake).

In order to improve the precision of food identification, barcode reading could be used for commercial foods. In some regions such as the European Union, each commercial food bears a barcode that can be read by a barcode reader. With the help of the barcodes, the food can be exactly identified, and information from nutrient labels can be directly used. However, it is still necessary to estimate the portion size of this intake in a proper way and to link the recorded data with reliable food composition data.

44.2.1.2 Standardization and Web-Based Approaches of Food Recording

Food records have been standardized and simplified by some groups (Koebnick et al. 2005). Paper versions were developed in which predefined categories of foods and portion sizes, often covering several days, are offered. The participant marks the respective items if the food has been consumed. Compared to the original method, this type of standardization reduces variability between the individuals in the study. The advantage of this simplification is the restriction to specific food items. The disadvantage is that it may miss foods that may later be considered important.

Computer-based versions of food records including their standardized and simplified versions can be found on the Internet. Often, they are directly linked to national food composition tables, among others, to allow nutritional counseling (e.g., Nutrition Analysis Tool (NAT) 2013). Users can enter their data using common web browsers. Such food records could also be used in the wider context of monitoring lifestyle changes in health programs, for example, weight loss programs. Computer-based versions of food records allow the rapid calculation of energy and nutrient composition and enable tailored guidance to be provided based in this information to foster behavior change. The aim in such circumstances is not so much accuracy as ease of use in affecting behavior change. Reliability and validation studies of these novel form of food records are rarely found in the international literature (Illner et al. 2012).

Major efforts have been undertaken at the beginning of this millennium to make use of new interactive devices such as mobile phones and personal digital assistants (PDAs)/tablets in order to record food intake (Illner et al. 2012). The advantage of these tools is that they allow direct and immediate reporting of food intake, even during a meal or directly after it. This method may not generate better results than paper-based food records; however, it has the advantage that food coding can be organized interactively using computer program-based analyses. As with the computer programs of recording diet from the Internet, the interest of those providing the mobile phone and PDA programs is often less research but more about dietary counseling and monitoring of dietary changes. This is well accomplished by an interactively communicated report of the energy and nutrient intake during recording. With the new generation of mobile phones, apps are available that ask to record food intake and calculate energy and nutrient intake.

44.2.1.3 Food Recording by Photographs

Within the US program, "Technology Assisted Dietary Assessment," mobile telephones of the new generation will facilitate a completely new way of collecting food records with the aim to obtain data with a much higher accuracy and validity than before (Daugherty et al. 2012). The idea behind this approach is not to record the meals but to produce images of all foods eaten. Photographs of meals on a plate can be taken before and after eating to allow subtraction of uneaten foods. The images are sent to a server of the institution and immediately analyzed by computer programs in terms of foods being used and the quantity ingested. It is the aim to train the computer programs to identify the foods properly and to calculate quantity even if the food is partly or mostly covered by other foods. Some companies and institutions have already established this method in a semiautomatic way in that dietitians assign foods and quantities from the pictures supported by computer programs. The semiautomatic approach is of commercial interest since the user may be willing to pay for the result. In the research environment, it is the aim to reduce the human factor as much as possible and to rely on powerful self-learning computer programs.

The current challenge is the quantification of foods. Initially, each image has to contain a standard that gave information about size and color. This might be a complication that makes the method less easy to conduct. Other ideas favor taking pictures from several predefined angles allowing producing three-dimensional images from which the quantity is derived.

With the progress in information technology and computer science, food recording by photographs seems to be the most promising novel approach to assess individual diet.

44.2.2 Food Recall

Unlike the food record techniques that record current intake, the food recall methods estimate past intake, sometimes referred to as a retrospective method. The most commonly used time frame for the recall is the intake at the past day, and as such, although it is a retrospective method, it is designed to cover current intake. The food recall can be done either by the respondent him/herself or by an interviewer. If only the past day is covered, the method is called a 24-h recall. A 24-h recall starts with an overview over the meals of the past day and works through each meal in detail by identifying each food being eaten including specifications such as brand name, cooking method, and fat content. Recall of the portion sizes may be aided by photographs and food models.

The main concern with this approach is that it relies on subjects accurately recalling what they ate. As with the record techniques, people may either deliberately or accidently exclude certain foods. Bias due to change in diet is thought to be less likely to occur if subjects are given no notice of the request for information. Otherwise, the 24-h recall will be as vulnerable to under-eating and social desirability as the food record. The risk of underreporting of foods when conducting 24-h recalls can be ameliorated by training of interviewers, use of standardized computer programs with meal orientation and tiers of questions with increasing details regarding the food being eaten, probing questions regarding in-between meals, and rechecks when a low-energy consumption has been reported.

The method of 24-h recalls is also considered similar to the method of food records to serve as a reference method for other less accurate methods. However, the retrospective and self-reporting nature of the 24-h recall can also induce bias and misreporting. As with the food record, about 10 to 15% of the consumed energy will not be reported, if compared with objectively measured energy expenditure (Poslusna et al. 2009).

Computer-based 24-h recalls can be linked to the latest national food and nutrient databases, where facets such as brand name, preparation, cooking method, and fat content can be incorporated for all foods or for specific food groups.

44.2.2.1 Interviewer-Administered 24-h Recalls

Traditionally, 24-h recalls are conducted by face-to-face interview, although telephone interviews rather than face-to-face are becoming more common particularly for large studies. Comparison studies did not reveal any statistically significant difference between face-to-face and telephone interviews (Tran et al. 2000; Brustad et al. 2003). The interviewer can use a computer program and ask about the diet of the previous day. Such an interview takes about 20 to 40 minutes. This type of 24-h recall has been applied in some of the calibration studies (see Sect. 44.5.2.) when a study participant was in a study center for examination. In some of these studies, home interviews were conducted if more than one 24-h recall was required per subject.

In order to obtain several 24-h recalls from a subject for randomly chosen nonconsecutive days, the telephone interview is the preferred method. In this case, the participant may receive a book with portion sizes at the beginning of the study, mostly during physical examination at a study center (and eventually coupled with a first face-to-face interview) and may then be waiting for the call by the interviewer on subsequent days.

44.2.2.2 Web-Based 24-h Recalls

Web-based 24-h recalls for self-administration have been developed and follow the same principles as the 24-h recall administered by interviewers. A good example of such an approach is the NUTRINET-SANTÉ Study, conducted in France (l’Étude Nutrinet-Santé 2013). The 24-h recall of this study is self-administered and web-based, developed in a long tradition of experience. In France, the first approaches to collect detailed dietary data by self-administration have been started in the early teletext times. Other web-based programs for self-administration have been developed by a Californian group (Arab et al. 2011) and by the NCI (Subar et al. 2012). The US programs have been organized similarly as the interviewer-based multiple-pass method, developed for the surveys of the US Department of Agriculture, and has shown to be feasible. A special tool was developed for adolescents in the HELENA study (Vereecken et al. 2008) which was adapted for young children in the framework of the IDEFICS study. This program is still computer-based but will soon be available as a web-based instrument (SACANA) for self-completion instrument (Börnhorst et al. 2013). However, it is still unclear whether there is a loss of information by the self-administration compared to interviewer-conducted 24-h recalls and how large the loss may be. Some loss of information by self-administration might be expected due to the complexity of current diet and skipping of probing questions. Therefore, the web-based versions of 24-h recalls cannot be regarded as quantitative reference methods yet.

A further simplification of 24-h recalls, designed for self-administration via Internet or touch screen, is done by keeping the meal-based structure but to offer only a selected number of foods and ask for their portion size and frequency of intake at that day (Biobank UK 2013). The time to fill in one questionnaire takes about 11.5 min on average. The questionnaire has been developed by an Oxford research group and is currently used in the UK-Biobank study (Liu et al. 2011).

Another modification of the 24-h recall has been done in Germany with the aim to have a self-administered instrument that focuses on the use of a selected number of foods on a specific day. This instrument has been designed to be applied as a web tool but can be also printed out and used as a paper version to be filled in at home (for the past day) after a telephone call (Boeing, personal communication; German Institute of Human Nutrition (2013) "Simplified 24-h-recall"). Filling-in time is about 8 min per single day. The simple self-administered 24-h recall allows repeated application without much time effort for the participant but needs further data on portion sizes from interviewer-conducted 24-h recalls and frequency per day information from a FFQ for those foods eaten daily.

44.2.3 Dietary History Method

The dietary history method goes back to Burke (1947) who proposed to use a habitual instrument to capture individual diet in addition to food records estimating current diet. The diet history is a retrospective method with the aim to obtain a detailed picture of the usual dietary intake of an individual. By this method, the individual has to describe usual food and supplement intake, organized according to meals, for a relatively long period, e.g., 4 weeks, 6 months, or 1 year. The method usually requires the administration by an interviewer. The face-to-face situation allows the use of common household measuring containers to estimate portion sizes. The method is popular particularly in the Scandinavian countries and has also been used in the health surveys of Germany, conducted by the Robert Koch Institute. The method has been the dietary assessment method of choice in a larger European study on the elderly (van Staveren et al. 1996).

Bälter from the Karolinska Institute, Sweden, has developed a web-based version of a diet history for a large-scale prospective study (Bälter et al. 2005). This web-based dietary history combines the meal orientation of food intake during a day and the focus of the food frequency (see Sect. 44.2.4) on habitual diet and is asking about habitual meal structured food intake.

As with other retrospective methods asking for habitual diet, social desirability and overestimation of healthy foods are one of the problems when compared with reference methods (see Sect. 44.5.1). Estimating habitual diet by the respondents themselves and not by statistical methods from short-term quantitative instruments (see Sect. 44.4) is usually prone to many biases due to the considerable intellectual effort of the participant to recall precisely the habitual diet during a certain time period.

44.2.4 Food Frequency Questionnaire (FFQ)

The method of the food frequency questionnaire (FFQ) became prominent with the first large-scale cohort studies, comprising more than 100,000 subjects, at the end of the 1970s. At that time, also the first data processing centers based on large computer units had been established, and it was logical to make use of the "new" technologies to collect data on dietary intake. For this purpose, the method of asking for the habitual eating frequency of portions of selected foods over a defined retrospective time period in a self-administrative manner was ideal and thus, the food frequency questionnaire became the leading dietary assessment method for all types of studies in which the diet of cases are intended to be compared with the diet of controls. This could be used prospective studies as well as retrospective case-control studies. The response categories in the first versions of the FFQ were initially confined to nine options, in order to optimize computer storage place. The FFQ is primarily a ranking instrument developed for comparing groups and less an instrument by which diet is exactly quantified.

The selection of foods to be included depends on the research question in general and the type of dietary exposure of interest. An FFQ designed for general purposes should be able to measure key dietary components related to health and energy intake. Key dietary components include foods and food groups as well as nutrients. The development of an FFQ requires a survey type of dietary data collection by reference methods in the source population in which the FFQ should be applied. Key foods, explaining most of the variance of the nutritional exposure, can be identified by methods such as stepwise regression and specific programs for that purpose such as Max-R and MOM (Mark et al. 1996). Also foods contributing to a specific nutritional exposure in high absolute amounts should be identified. Both types of considerations, variability of food intake in the source population and the contribution of foods to the nutrient and energy intake, finally result in the list of foods selected to be included in the FFQ. Usually, the number of food items in an FFQ is about 130 to 180. Each FFQ needs investigations in as much the FFQ is able to reflect true dietary intake (see Sect. 44.5.1) and whether calibration strategies (see Sect. 44.5.2.) could be applied to assure that the dietary intake values reflect the intake in the source population (Nöthlings et al. 2007).

The focus on food items makes the FFQ very attractive to compare subjects and populations regarding the pattern of dietary intake. Some FFQs have already been developed to be applied in a multicultural setting, for example, in the multiethnic cohort of Hawaii. This study consists of subjects of Hawaiian, Japanese, Caucasian, and black origin (Kolonel et al. 2000). Also, an FFQ was developed for estimating diet in different European countries (Illner et al. 2011). In view of worldwide trends toward a unified food supply which is usually accompanied by a reduction of specific cuisine traditions, multicultural FFQs asking for the use of commonly available foods as well as some specific traditional foods in all study subjects might be feasible and give a first insight into still existing differences in eating habits of subjects and groups on a wider scale based on one instrument.

44.2.4.1 Traditional FFQ

From its beginning, the paper version of an FFQ was designed for optical scanning. Often, the requirements for optical scanning dominated the design of the instrument, for example, that the frequency scale for each food item was displayed in lines. The number of lines per page, in turn, followed the number of sensors of the scanner.

In the FFQ usually the frequency of intake of a standard portion over a year is requested. For case-control studies, the time period to be covered is usually extended, for example, to 5 years before onset of the disease or the last 5 years. Seasonal variations and specifications regarding facets of the food such as fat content (dairy foods) were often included into the FFQ. Further research revealed that the frequency scale should be vertically organized instead of horizontally. Most individuals preferred this type of presentation in a feasibility study (Thompson et al. 2002). It is still a question as how to incorporate the portion sizes into an FFQ. Since the frequency scale always makes reference to a portion, the assessment of the individual portion size improves the ranking of a subject and increases the comparability between subjects. Thus, in traditional FFQs, questions on the portion size of a food item, categorized in small, medium, large – often illustrated by pictures – are asked. Portion size information usually accounts for 10 to 20% of the variation in food intake (Noethlings et al. 2003). It is still the question whether such portion size information is reflecting true behavior. Research in recollecting and proper assignment of portion sizes revealed that portion size information has some validity (Haraldsdóttir et al. 1994; Ovaskainen et al. 2008).

44.2.4.2 Food Propensity Questionnaire (FPQ)

In connection with considerations of combining information of 24-h recalls with FFQ information (see Sect. 44.4), it was suggested to reduce the information requested in the FFQ to the frequency scale and to omit the questions on portion sizes (Subar et al. 2006). This step would reduce the burden for the study participant but could still make use of most of the information a FFQ is providing.

44.2.4.3 Screeners for Dietary Exposure

The selection of foods for the FFQ can be limited to only a few or even one dietary exposure. The resulting questionnaires are called "screeners" since they only screen and rank subjects according these (this) exposure(s) (Thompson et al. 2004). Screeners are particularly interesting if in a study only one or a few hypotheses are of interest and time and resources are limited to collect dietary data. However, since the screeners lack the information on energy intake, adjustment for energy intake allowing to control for the different energy needs of the study subjects and to improve the validity of the estimate of the dietary exposure is not possible (see Sect. 44.5.2).

44.2.4.4 Web-Based Versions of the FFQ, FPQ, and Screeners for Dietary Exposure

From the beginning, the use of FFQs as a self-administered instrument facilitates its application as a web-based tool (García-Segovia et al. 2011). With the Internet, study participants can now fill in the questionnaire at home on the computer or directly in the study center. In the study center, the questionnaire can be offered as a touch screen program. Studies comparing the traditional paper versions with the Internet versions did not find differences in the performance of the instrument. Many research groups now offer paper as well as Internet versions of their FFQs.

44.2.5 Nutritional Biomarkers

Nutritional exposure can also be expressed as nutritional status defined through biomarker measurements. Nutritional status reflects the dynamic balance between dietary supply, body pools, and the metabolic demands. Measuring food intake does not represent a nutritional status. At the same level of dietary intake, a subject may be able to function well or poorly, depending on the demands being placed upon him/her. If an individual is growing or fighting an infection, the dietary supply of a nutrient may or may not be adequate to enable and maintain optimal function. However, the available nutritional substrates for function at a certain time period not only are coming from the diet but can also been mobilized from the body pools. Where there is competition for these metabolic substrates, one function may be compromised by another one. In an epidemiological study, it is often assumed that a reported measure of intake reflects the functional availability. However, this assumption is rarely true. Nutritional biomarkers are primarily obtained from blood or urine to estimate nutrient status and/or to validate nutrient intake obtained by (self-reported) dietary assessment.

In this chapter, we will not address assessing nutritional status by biomarkers or the development of appropriate methods to identify new biomarkers of nutritional status. Novel -omics techniques allow a completely new and comprehensive approach with simultaneous determination of thousands of metabolites. Thus, biomarker studies with this type of data will play a prominent role in the future regarding the search for a better characterization of a subject regarding nutritional status. Here, we address which biomarkers are useful to validate dietary assessment instruments. Biomarkers have a major advantage of being independent from all self-reported diet with respect to their error structure. This is a major statistical advantage when statistical models are built to identify measurement error (see Sect. 44.5.1).

Nutritional biomarkers can be divided into three categories: (1) recovery biomarkers which are considered to capture true dietary intake and are measured in the same units as dietary intake by dietary assessment methods such as nitrogen in urine, (2) concentration biomarkers which correspond to dietary intake but do not reflect total dietary intake and are not measured in the same units as dietary intake such as vitamin C, and (3) replacement biomarkers that should be used instead of measuring intake by dietary assessment, i.e., when it is difficult to capture dietary intake due to limited information in nutrient databases, widespread occurrence in foods, or very difficult measurement procedures by questionnaires such as selenium.

For validation purposes, the recovery biomarkers are the primary choice. However, the number of true recovery biomarkers is small and refers to energy expenditure and nitrogen excretion in urine, followed by the minerals potassium and sodium, also excreted in urine. With the exception of energy expenditure measured by doubly labeled water, urine is the biomaterial of choice for recovery biomarkers. Less suitable than recovery biomarkers for validation purposes are concentration biomarkers. They are usually determined in blood such as vitamins and bioactive compounds, e.g., vitamin C, folic acid, vitamin D, or carotenoids. Concentration biomarkers usually lack a linear relationship with intake over the whole range of exposure, and their concentrations in blood are the result not only of intake but of metabolic activities including genetic makeup.

Urine samples for the determination of recovery biomarkers require 24-h collections. The collection of complete 24-h urines is laborious for participants and researchers since it requires the permanent availability of collection containers with preservatives and their storage and a pickup service. Some researchers also favor the use of 4-aminobenzoic acid (PABA) or other markers to verify the completeness of urine collections (Murakami et al. 2008). With PABA as a standard, correction factors for incompletely collected urines can be applied.

44.3 Defining Dietary Exposures

In a research setting, the hypothesis is framed in terms of how the exposure influences the outcome. In this sense, the measure of dietary exposure needs to reflect the relevant exposure in the formulated hypothesis. Nutritional status refers to measures of diet, body composition, and biochemical and clinical states. All can be used as measures of exposure. In this section, we characterize dietary exposure in more detail. Diet is a general term and comprises all that we eat. Food describes the individual items that make up a diet. The dietary assessment methods record these individual items (and also supplement intake). Food intake can be converted into intake of nutrients, other food compounds and energy by multiplying it with the average concentration of these compounds in the food. In rare occasions food intake could also be duplicated and one half ingested and the other half chemically analyzed for particular constituents, including, e.g., products formed by different cooking methods. Intake of individual foods, nutrients or dietary patterns may be the relevant exposure, depending on the hypothesis being tested.

Food composition tables are the means to convert food intake to nutrient intake. A normal supermarket can offer several thousand different food items and the overall number of available foods in a country approaches 200,000. Food-composition tables are usually hierarchically organized from major food groups to subgroups down to single food items. Food composition tables usually do not contain brand names but are structured according to foods. They are food-oriented and thus contain information on the processing, e.g., raw, cooked, pickled, and canned. They also include ingredients and foods we never eat as described. Complex dishes and recipes pose a problem for food coding. Sometimes, the conversion into ingredients is done, and sometimes a general food code is used. Thus, it is important to follow in each study the same strategy in order to make dietary assessments comparable. The different strategies of food coding and the use of different food composition tables also make it difficult to compare data from different countries directly.

Food composition tables link the nutrient content or the content of other food constituents with a given food item. The contents or concentrations of the constituents are often generated as a combined effort of several institutions and laboratories (Bell et al. 2012). Application of standardized laboratory methods and quality control measures are essential to compile meaningful food composition tables. Often, an analytical method is changing with technical progress and laboratory data nowadays are not directly comparable with older data on food constituents. A particular problem for researchers could be the variation of the constituents within a food. Species differences, ripening status, and soil condition are a few of the factors that could increase the variation of the constituents. An example is selenium in grain which heavily depends on soil concentration which is quite different in the USA compared to Europe. Selenium concentration in grain therefore depends on the annual import and export figures between the USA and Europe. If the variation of a given constituent in a food is very high and if this constituent is contained in many foods, one might conclude that the use of biomarkers might be a better choice to rank subjects.

44.3.1 Defining Reference Categories and Cut-Points

In statistical models, nutrient or food intake data are commonly used as categorical variables rather than as continuous ones (see also Sect. 44.6). Categorical variables are created by dividing the continuous data of the study population into tertiles, quartiles, or quintiles. The categorical approach makes no assumptions about the underlying shape of the association between the nutrient and the outcome. Even if intakes are used as continuous variables, analyses based on categories should be done to check whether the assumptions made for the continuous approach are justified. For example, in logistic regression, the assumption of a linear trend can be assessed by examining the variable in categories (Hosmer and Lemeshow 1989; cf. chapter ▸Regression Methods for Epidemiological Analysis of this handbook). One reason in favor of the categorical approach is that it expresses the results for dietary factors in the same way, independent from the instrument being used and the random and systematic biases in estimating intake. Categories describe the measured ranges of intakes among the individuals in the study group in an identical manner. The relative risk estimates for the highest versus the lowest category can be directly compared to assess the relative impact on the outcome of interest. Nevertheless, caution may be required in comparing results from different studies where the range of intakes covered in the top or bottom third, for example, may be quite different. For example, comparing tertiles of intake of red meat between Germany and China, the intake in the highest third in China will probably be included in the bottom third for Germany.

Using external criteria might be an alternative to build categories. For example, one could use published dietary recommendations such as 30% of energy from fat or 30 grams of dietary fiber per day. The use of external criteria, however, requires that the dietary estimates in the study are not biased. Otherwise, the public health message coming out of the study might be grossly misleading. The same applies to the labeling of the categories. The values used to label categories should be meaningful to the readers. Otherwise, it would be better to label categories according to an ordinal scale.

Where data are recorded and reported as continuous variables any associated disease risk is expressed per unit of exposure. Thus, it is important to assess the exposure in terms of absolute amounts consumed. The respective unit of exposure can then be selected to express the change in relative risk. For some exposures, units of 1 g are biologically meaningful, while for others units of 10 g or 100 g are of relevance. The exposure units can also be chosen according to external knowledge, for example, from dietary recommendations.

44.3.2 Measurement of Dietary Behavior

Dietary behavior can be an important aspect of dietary exposure. Dietary behavior can be asked directly in questionnaires. Corresponding questions address, for example, for cooking methods, food preparation, food sources, or food knowledge. For health outcomes, food preparation and in particular cooking methods may be important. In this respect, meat preparation has gained the highest interest so far. Some groups such as from the NCI developed and validated specific questionnaires with pictures about the way a subject is preparing and eating meat in order to estimate the intake of carcinogenic compounds (Sinha 2002).

The long-term observation of dietary behavior in countries that are in transition or in specific groups that undergo major cultural changes seems to be of high scientific interest in order to understand the impact of changes in dietary behavior on health outcomes. In these instances, the consumption of commercial foods, eating out of home, and sources of food supply are repeatedly assessed and analyzed in the population of interest.

44.3.3 Dietary Pattern Analysis

There has been a growing interest in approaches to summarize the dietary data by forming patterns and by investigating their association with disease risk. Dietary patterns can be assessed for foods as well as for nutrients. Most of the dietary pattern analyses are still devoted to foods. Dietary patterns can be derived in different ways (Hu 2002). One method is to use some a priori understanding of the way in which people should eat. Based on this understanding, indices are formed which describe how well an individual is following the prescribed dietary pattern.

Examples for such indices are the Healthy Eating Indices developed for the USA, the Mediterranean Diet Index, or the Nordic Healthy Eating Index. Other ways of forming dietary patterns are based on exploratory statistical methods such as principal component analysis, reduced rank regression, and cluster analysis. The latter methods make use of the existing empirical data collected in a study and utilize the correlation or covariance matrix between the dietary exposures. They are also called a posteriori methods in contrast to the a priori methods.

44.4 Methods to Calculate Habitual Diet Based on Short-Term Food Records and Recalls

For a long time, it was discussed how data from 24-h recalls and food records can best be incorporated into studies aiming at habitual diet. Some researchers argued that time periods of collecting data of only a few days are not sufficient to characterize the habitual diet of individuals. They highlighted that day-to-day variation and seasonal intake might make such methods less suitable for epidemiological studies that relate diet with disease risk or other personal characteristics. Nowadays, it seems possible to collect several 24-h recalls per individual at random over a year, for example, by telephone (see Sect. 44.2.2). Further, statisticians have developed complex programs that allow an estimate of the habitual diet based on these 24-h recalls. In this section, the aim is to describe how to calculate habitual diet from short-term instruments with their major research lines.

Dietary data, derived from 24-h recalls and food records, include two types of variation. One type of variation is the daily variation of dietary exposure in an individual (within-person or intraindividual variation) and the other type of variation is the variation of dietary exposure between the individuals (between-person or interindividual variation). The total variation of dietary exposure data of a study is therefore the sum of the intra- and interindividual variance. In a study with many study subjects, the interindividual variation is usually causing the major part of the variation. The extent to which the intraindividual variation contributes to the total variation of the dietary exposure depends on the number of days recorded for an individual. Each individual has a true mean value per day around which the single-day values vary. With increasing days recorded per individual, the estimate of the true mean of the individual becomes more precise. If all 365 days of an individual would have been recorded unbiased, the true mean intake per day of all individuals can be defined precisely. With fewer days, the individual mean values still vary around the true mean.

The existence of the two types of variation has several consequences. One consequence is that the distribution of the mean intake in a study population does not reflect the true distribution of the means among the individuals. This may be important when public health measures are derived from such distributions (see Sect. 44.7). The deviation from the true mean of an individual increases when the number of days per individual decreases. The distribution of dietary intake data with only one recorded day per individual is usually much flatter than the true distribution due to the additional intraindividual variation. With more recorded days per individual, the estimated distribution in the population becomes more narrow. Another consequence is the dependency between the ratio of the two types of variation and the number of days needed to characterize an individual regarding a specific dietary exposure. The more stable the daily dietary intake is, the smaller is the intraindividual variation, and the fewer days are required to characterize the dietary intake of an individual. Protein, for example, is a dietary constituent with only a comparatively small day-to-day variation. On the contrary vitamin A has a high intraindividual variation due to occasional intake of vitamin A-rich foods. For example, liver is rarely eaten (see Sect. 44.2.1.1).

The implication of the different types of variation within a dietary data collection with 24-h recalls or food records on the practice of nutritional epidemiology was intensively discussed in the 1980s (Beaton et al. 1979, 1983). It was proposed to use variance component analysis to calculate the two types of variation and to remove the intraindividual part of the variation in order to obtain the true variation between the individuals (Beaton et al. 1979, 1983). Traditional variance component analysis requires a normal distribution of the intake data which is only rarely found. Perhaps this is why this procedure does not improve the estimate of dietary exposure for the individual but does so for the whole study group.

Progress has been made during recent years to develop the statistical framework further in order to identify and control for the intraindividual variation properly. The aim of such statistical procedures is to estimate the true study distribution based on only the interindividual variation. The estimated true study distribution also provides the best estimate for each individual. An introduction to the statistical background of such methods can be found on the NCI website (National Cancer Institute 2013) as well as links to their statistical program. The statistical method transforms any type of intake distribution into a normal distribution, removes the intraindividual variation by variance component analysis, and back-transforms the new values to the original scale. Further, the analysis has been split into the part of estimating the probability of intake (i.e., of a day) and of the amounts eaten at each occasion.

Further efforts try to extend the approach to the multi-exposure situation, since one exposure correlates with other exposures in terms of error structure and variance components (Zhang et al. 2011). Such statistical tools for the multi-exposure situation, particularly if energy is included (see Sect. 44.6.1), will result in even better estimates of the distributions compared to only considering one exposure at a time.

The use of such methods is straightforward when nutrient data are investigated. Nutrients are consumed almost always every day, and non-zero consumption is not an issue. If foods are investigated, non-consumption might take place at the day on which the data are collected. Even if the recall spans several days, no consumption of a food might be recorded despite the individual has eaten the food outside the recorded time period. Thus, it is proposed to make use of the frequency information of an FFQ if such data are available. The FFQ provides information of the habitual use or non-use of a food. A habitual non-user of a food should have a zero intake. In contrast, a habitual user of a food even if no data are available on the 24-h recalls or food records should have an intake value even if it is small. Some of the statistical programs take this into account and simulate the intake of foods not reported in 24-h recalls.

The estimates for each individual, calculated by such programs, can be regarded as the best estimate of the habitual diet derived from the short-term dietary assessment instruments (George et al. 2012). These estimates are not identical with the mean intake over the days which are usually calculated for each individual from food records or 24-h recalls. The estimation of the habitual intake of an individual can be further improved by considering covariate information such as age, sex, BMI, and habitual frequency information of a food. It has been calculated that 4 to 6 days per individual might be sufficient to capture most of the food and nutrient intake of an individual correctly (Carroll et al. 2012). It has also been found that often the individual estimates improved when the FFQ information is added (Carroll et al. 2012).

In view of the advantage of short-term methods such as 24-h recall and food record with regard to their precision and quantity, their use to calculate habitual intake is considered an important step further. The new statistical programs might facilitate up the widespread use of quantitative assessment methods in future epidemiological studies.

44.5 Quantification and Correction of Measurement Error

Different study designs use different methods to assess diet, often for reasons of cost, time, and effort. However, it should be the aim of nutritional epidemiology to assess and utilize dietary data that are close to true intake (Fig. 44.2).

Therefore, a great deal of research and effort has gone into the validation or calibration of methods used in nutritional epidemiological studies (Table 44.1). The aim of such studies is less to improve the method itself but to define, and then correct for, the type and amount of measurement error (calibration studies). The validation of a method is usually done by comparing the empirical data from the dietary assessment with a measure of true intake. However, true intake (truth) is nearly impossible to measure. True intake should be considered more as a construct or latent variable than the result of a direct measurement. As a compromise, researchers take the best possible methods, usually 24-h recalls or food records, and compare their results with the results from the study instruments such as FFQs, diet histories, and screeners. This type of validation is also called relative validation (or some prefer to comparative study), since the reference method still contains measurement error compared to the truth. The reference method is often labeled as the relative reference method compared to a gold standard reference method. In this comparison, it is assumed that the errors of each method are not correlated and that any agreement observed between both measurements derived is not due to this correlation of errors.

Including a biomarker may provide further refinement to the validation of dietary methods. It is assumed that the error associated with the collection of biomarker information is independent from the errors of self-reports of dietary intake. This assumption allows the use of complex statistical methods to estimate true intake. This framework is also known as the method of triads (Daurès et al. 2000). In the method of triads, the correlations between questionnaire data, reference data, and

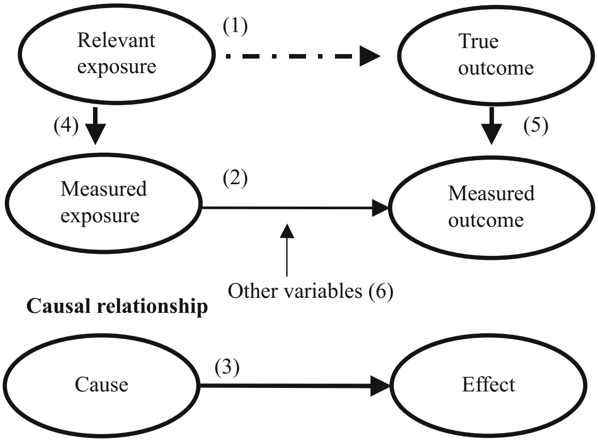


Fig. 44.2 Relationship between exposure and outcome, cause and effect. (1) This represents the true relationship between relevant exposure and outcome. (2) This represents the observed relationship between measured exposure and measured outcome. (3) This represents the true causal pathway. The cause must precede the effect. (1) and (3) are the same if other variables (6) (confounders) are either absent (no residual confounding) or taken account of (by stratification or statistical adjustment). (4) This represents the relationship between relevant exposure and measured exposure. (5) This represents the relationship between true outcome and measured outcome. (6) Variables that should be measured and reported that may influence the relationship between exposure and outcome.
The extent to which the measured exposure and outcome deviates from the true and relevant measures should be described in a validation study that presents the measurement errors

biomarker data are used. In its most advanced form, the statistical framework of the method of triads includes repeated measurements of all three assessment methods (Rosner et al. 2008).

44.5.1 Quantification of Measurement Error

The way that measurement error is assessed needs to be appropriate as to how the data derived from the method will be used in the main study. It should also be clear that the sample of people used to assess the measurement error should be similar to the study population in terms of characteristics. In broad terms, the comparison study assumes one method is a better estimate of the true exposure (reference method) and is often denoted by Q, while the test method is denoted by R. Ideally Q equals R in any comparison. There are a number of different ways to assess and express the relationship between R and Q. For example, the relation between R and Q can be described by correlation coefficients. Validation studies also established cross-tables between R and Q and described the relation between R and Q with a categorical association measure such as kappa or chi-square. The decision of the way to describe the validity of a method depends on the statistical approach of the

Table 44.2 Probabilities of misclassification of a reference ranking in fifths using an imperfect alternative that has various correlations with the reference (From Walker and Blettner 1985)

Absolute difference in quintile ranks	Correlation coefficient between the reference and alternative			
	0.9	0.7	0.5	0.3
0	0.573	0.403	0.321	0.263
1	0.378	0.400	0.379	0.355
2	0.047	0.156	0.203	0.225
3	0.002	0.037	0.081	0.118
4	0.000	0.003	0.017	0.038

main study. Alternatively, the validity can be described by using several approaches in parallel to generate various types of validity coefficients. If the study aim is to compare the relative risk of one level of exposure with the relative risk of another level, then the validation study should also assess how well the selected method can distinguish these levels compared to the reference method.

A regression approach allows a more sophisticated analysis of the relationship between Q and R.

The respective relation can be described with

$$R = \alpha + \lambda Q + \varepsilon,$$

where α describes the systematic bias, the parameter λ the ability of Q to measure the range of intake in R, and ε describes the random error. A small value of λ indicates a low ability of the instrument Q to assess existing differences between study subjects. The estimates of λ are sometimes used to correct the observed relative risks for measurement error (Spiegelman et al. 1997). It is difficult to get measures of α, and it is often assumed to be zero. If this assumption is not true, then a distorted estimate of agreement will be derived.

Table 44.2 shows the proportion of people who are classified into the correct quintile by the reference method but into a different quintile by the alternative/test method (Walker and Blettner 1985). For a measure with a correlation of 0.5, only 32.1% of individuals are placed into the correct quintile, another 37.9% are one quintile too high or low, while the remaining individuals are misclassified by two quintiles or more. This table also shows that extreme misplacement over four quintiles is rare. The impact of the measurement error on the study result is profound. Assuming a cohort study with a "true" incidence of 2% in the lowest quintile and 6% in the highest (i.e., a relative risk of 3.0), using the alternative method with a correlation of 0.5 with the reference method would lead to incidences of 3.1% and 4.9% in the lowest and highest quintile, respectively, and result in a relative risk of 1.58 instead of 3.0 (Walker and Blettner 1985).

A particular concern is whether those who misreport their intake are different in other important characteristics from those who do not. Most work has been done on misreporting energy intake by individuals with different levels of body mass index.

The data suggest that overweight people tend to underreport their fat intake. Another issue is misreporting due to social desirability and established norms. Recent work showed that fruit intake is overreported by low consumers because they know they should eat a certain amount of fruit (Amanatidis et al. 2001). This phenomenon was particularly observed among more educated individuals. Before a study begins, it may be possible to take a tendency to differential over- or underreporting into account in the design of the study, for example, by excluding overweight people. However, if this is done, it could mean that the study results cannot be generalized to the whole population.

Even if a validity study has not been conducted, the test-retest correlation (reliability) (cf. chapter ▸Epidemiological Field Work in Population-Based Studies of this handbook) can provide an indication for the maximum level of validity. The correlation between the test and reference measures cannot be greater than the correlation of repeat measures of the test measure (Walker and Blettner 1985). Low test-retest reliability clearly indicates also low validity.

This initial concept of validation with a reference measure was extended toward measurement models that included objective measurements such as biomarkers. Biomarkers can be used as the only reference method or as a second reference method. The use of biomarker information as a second reference method leads to the method of triads and latent variables (see Sect. 44.5). A second reference method gives also a sense of how well the first reference method may be in measuring the underlying truth.

Some of the biomarker information can be considered to provide gold standard information on the nutritional exposure of interest (see Sect. 44.2.5). In contrast to the relative reference methods, the gold standard information provides a measure of nearly the truth. However, we still have to be aware that this term gives a false sense of accuracy due to the often short-term recording periods of the gold standard reference. One of the gold standard reference methods is the measurement of energy expenditure as a proxy for energy intake by doubly labeled water (DLW). DLW provides an integrated measure of components of energy expenditure, usually over 7 to 14 days. The technique involves ingestion of a small amount of water with hydrogen replaced by deuterium, where both the deuterium and the oxygen are isotopically labeled. As water molecules are split during energy metabolism, oxygen binds to carbon and is exhaled as CO_2. The excretion of deuterium is then measured in urine. The rate of loss of these labeled elements can then be used to provide an estimate of energy expenditure (Goran and Astrup 2002). Apart from DLW, urinary nitrogen is also considered as a gold standard method. Aside the two, there are only a few other robust independent measures that are relevant (see Sect. 44.2.5). Most of the biomarkers of nutritional status lack a linear dose-response relationship with the level of dietary intake (which is the relevant exposure) (Bates 1997a, b). If there is a curvilinear relationship, the question is important how the intake of the study population fits into this relation. If the intake of the study population is either at the lower or higher end of the distribution, the comparison of the dietary and the blood measure may show poor agreement and lead to the erroneous conclusion that the dietary assessment is poor.

Studies with DLW seem to suggest that there is no or only a low relation between energy intake assessed by an FFQ and energy expenditure derived from DLW, particularly in overweight or obese subjects. It seems that, based on DLW, many overweight people underreport their energy intake (Subar et al. 2003). A further refinement is to express energy intake as a function of the resting or basal metabolic rate (BMR). If energy intake is less than the BMR, it would seem reasonable that subjects who are not losing weight underreport their dietary intake. Usually a factor is added to the BMR to allow for some level of physical activity (PAL). A PAL value of 1.2–1.5, reflecting low to moderate physical activity, is often used. If multiplied with BMR, the estimated required energy intake is obtained. In the national diet and nutrition surveys of British adults, using a 7-day weighed food record, about 40% of the subjects underreport their intake, assuming some degree of moderate activity (Black et al. 1991). However, underreporting has been found across all PAL. Thus, underreporting is not completely compensated when individuals with an energy intake below a certain BMR to PAL level are excluded from the study (Tooze et al. 2012). In case of the diet and nutrition survey of British adults, excluding 40% of subjects due to underreporting energy intake reduces the power of the study and may cause selection bias but may be necessary to reduce information bias.

One of the important studies with biomarker information was the Observing Protein and Energy Nutrition (OPEN) Study conducted in the USA from 1999 to 2000 due to its design and size. This study assessed the dietary measurement error of two self-reported instruments: a semiquantitative FFQ and a quantitative 24-h recall (Subar et al. 2003). Doubly labeled water and urinary nitrogen were used as biomarkers of energy and protein intake in nearly 500 men and women. In this study, the self-administered dietary assessment instruments were only able to pick up a minor part of the true variation between individuals as compared to the objective measures. The FFQ used in this study reflected 4–16% of the true variation in energy intake and 30–40% of the variation in protein density. The four 24-h recalls reflected 20–37% of variation in true energy intake and 35–50% of the variation in protein density.

44.5.2 Correction of Measurement by Calibration of Dietary Data and Measurement Error Models

The main idea behind calibration of the dietary data in a study is to improve the estimates of dietary exposure of all study subjects by using data from a validation substudy collected within the main study. A validation study in a subgroup of the main study should be regarded as a general advantage independent from using the results for calibration since validity is assessed under real study conditions. The population of the sub-study should be representative for the study population, and the collection of the additional data should follow the rules of the method of triads.

Calibration means that the information obtained by the dietary instrument of the main study is corrected using the information from the validation study with the better method in a measurement error model. Usually, a weak dietary assessment leads to a loss of power and the attenuation of the risk estimate toward the null value. Calibration should reduce the attenuation of the relative risk. The concept of calibration is closely related to the concept of validation. Simulation studies could show that calibration in most instances fulfill the purpose to improve the estimate of the diet-disease relationship (Freedman et al. 2011).

There are several ways to establish the measurement error model. An often used approach is regression calibration, in which the relation between the self-reported diet and the biomarker information is described by regression equations (Freedman et al. 2011). These equations include regression terms for the characteristics of the participants that influence the estimate of the observed dietary data compared to the truth. For all subjects that are not part of the sub-study, an estimate of the true intake is calculated given the self-reported intake and the personal characteristics of that subject.

44.6 Statistical Approaches in Nutritional Epidemiology

In principle, the statistical risk modeling of dietary data is not different from the modeling of other data. They are usually used as continuous or categorical data (see Sect. 44.3) in Cox proportional hazard or logistic regression models.

In the past, most of the nutritional epidemiological studies presented relative risks for ordered categories such as quartiles or quintiles. Usually, the lowest category was selected as reference category, and the relative risk was estimated for the other categories of intake. Since the FFQ was considered as a ranking instrument, the use of categories seemed to be appropriate. The use of intake categories instead of a continuous measure also allows investigating non-linear relationships and thresholds (see Sect. 44.3). With more precise data, there is increasing interest to describe the relation between dietary intake and risk of a disease in a continuous manner. This trend is further supported by the estimate of habitual intake from short-term reference instruments (see Sect. 44.4) and the use of calibration and measurement error models with objective measures (see Sect. 44.5.2).

Thus, this section concentrates on those concepts that cover analytical approaches based on continuous data.

44.6.1 Adjustment for Confounding by Energy Intake

The requirement of energy depends on individual factors such as growth in children, body weight, and physical activity. The nutrients that provide energy include fat, carbohydrates, protein, and alcohol. There is high interest on the role of these energy-providing nutrients, particularly of fat, for disease risk. However, the amount of intake of energy and the intake of the energy-providing nutrients is highly correlated, and it is necessary to separate the effect of energy from the effect of the nutrient providing that energy. Further, with increasing energy intake, the proportions of the energy-providing nutrients change. With increasing energy intake, more fat is usually consumed due to its higher-energy content per unit of intake compared to other nutrients. Thus, energy intake is a confounder for the energy-providing nutrients. In respect to disease risk, energy intake itself does not seem to be an important exposure to be studied. Energy intake is a direct consequence of other important lifestyle factors for disease risk such as body weight and physical activity, which should thus be taken into account in the modeling strategy.

For many years, it was believed that calculating nutrient density would protect against the phenomenon of confounding by energy. Nutrient density means that the nutrient content of a food is put in relation to the energy content of a food by dividing units of nutrient by units of energy. However, energy is still indirectly included in the nutrient density term but as denominator. If for any reason energy intake is associated with the disease of interest – e.g., due to inadequate control of energy related risk factors – residual confounding by energy intake remains when using nutrient density. The only protection against this type of bias is the inclusion of energy intake in the statistical model in addition to the nutrient density term.

In 1986, Willett and Stampfer (1986) proposed to overcome the problem of confounding of the energy-providing nutrients through energy by calculating energy-adjusted nutrients and to use this term in the statistical modeling instead of the nutrient itself. Energy-adjusted nutrients were obtained by taking the residuals from a regression of energy to the energy-providing nutrient. The residuals of this regression reflect the amount of the nutrient which is consumed by the individual compared to the average consumed at a specific amount of energy. Energy adjustment by the residual method removes the variation due to energy intake from the variation of the nutrient. As a consequence, the original variance of the nutrient is reduced compared to the variance of the nutrient without considering energy intake. This will affect the observed range of nutrient intake in the study population. This range will be much smaller for energy-adjusted values compared to the original values. Residuals, however, have a mean of zero and contain negative values if individual consumption was less than the average. In order to obtain interpretable energy-adjusted nutrient data, a constant should be added to the residual, usually the mean or median intake.

Kipnis et al. (1993) showed that identical beta-coefficients should be obtained in a logistic regression for a nutrient when a standard model is applied by using energy and nutrient intake in the same regression equation compared to using the energy-adjusted nutrient. A different beta-coefficient for a nutrient is obtained in a logistic regression with a nutrient but without energy adjustment. The beta-coefficient of a nutrient in a logistic model with adjustment for energy reflects the relative risk which is obtained when a nutrient is substituted for "other" nutrients (energy is kept constant) and the beta-coefficient of a nutrient without energy in the model represents the relative risk when a nutrient is added to the diet.

In the measurement error models derived from studies with gold standard measurements such as in the OPEN study, often low correlations between measured and true energy intake were observed (Subar et al. 2003). Thus, without energy adjustment also the nutrient will be affected by this type of measurement error. It is therefore not surprising that in validation studies energy-adjusted dietary exposure variables show better correlations than the original variables (Subar et al. 2001).

Including energy in a regression model results in a model in which the energy is kept constant. Thus, increase or decrease of a dietary variable containing energy (fat, carbohydrates, protein, and alcohol) can only be done at the expense of another dietary variable also containing energy. Having energy in the regression equation with disease risk as outcome would result in investigating an isocaloric situation. Such a design is comparable with isocaloric intervention studies, in which one dietary component is exchanged with another dietary component in energy equivalents. Since the energy variable is in the ideal situation a linear combination of all nutritional variables contributing energy, not all energy-providing variables and energy could be in the regression equation. The variable which is exchanged needs to be left out of the regression equation. It is also of advantage to express each nutrient variable in energy units such as percent energy. The nutrient variable which is usually left out of the regression equation is the variable contributing most of the energy, often the carbohydrates. Isocaloric modeling is also called performing a substitution model. One energy-providing nutrient variable is substituted by another energy-providing nutrient variable in the risk model. Substitution models can also been run with other parameters being held constant. For example, if the overall fat content is held constant, one type of fat such as saturated fat could be substituted by another type of fat such as polyunsaturated fat. Other substitution models can refer to the amount of food. If the overall amount of food is included in a regression model with single foods, change in one food can only been done at the expense of other foods not being in the regression equation but contributing to the overall amount of food.

Some researchers have investigated substitution models for risk of a disease. For example, Daniel et al. (2011) showed that increasing poultry intake does not affect risk of colorectal cancer but decreases the risk if poultry replaces meat. Meat was associated with increased risk in that study.

44.6.2 Linear and Non-Linear Relationships

In their simplest form, statistical models assume and describe a linear relationship when continuous data are used. It is important to be aware of the fact that the models calculate the change in relative risk per unit change of dietary exposure. If the units are small, the relative risk unit will also be small. Thus, it is useful to rescale the original units into meaningful units, based on the level of consumption and feasible changes in consumption in the target population. As a guiding rule, a 10% change in dietary behavior would be considered as reasonable and worthwhile. Thus, for example, a change in fat intake from 100 to 90 g per day would be useful to know, whereas to assess a change in risk from 100 to 10 g per day would not, as this degree of change is hardly ever achievable.

There are many instances in which the assumption of a linear model can be wrong. First hints of such a situation can be obtained by using categories. However, in principle, it would be better to test whether a deviation from the linear model exists. The easiest way to test whether a non-linear relationship exists is the use of power terms for the variable of interest or of the spline technique. Power terms higher than the linear term can be tested for significance. There are several ways to apply the spline technique. The most commonly used method in the dietary field is to model restricted cubic splines. This means that simultaneously linear, quadratic, and cubic terms of the dietary exposure are modeled, and the extreme ends of the distribution are restricted to be linear. In addition, the exposure needs to be divided into sections by the definition of knots (cut-points).

A further technique to investigate non-linear relationships is by using fractional polynomials (Royston and Sauerbrei 2008). In a simulation study, the fractional polynomial technique seems to perform even better than the cubic spline approach to detect non-linear relationships (Binder et al. 2013).

44.7 Organization and Presentation of Data: Implications for Meta-Analyses and Reviews

We distinguish meta-analyses based on literature (MAL) from meta-analyses using pooled data (MAP). MAL is based on aggregated data taken from published studies, while MAP uses the raw data from individual studies and combines (pools) them into one single dataset. Both approaches to summarize the evidence help to understand the role of nutritional exposures for the occurrence of diseases and to plan potential prevention measures (cf. chapter ▸Meta-Analysis in Epidemiology of this handbook). The organization and analysis of published data as quantitative meta-analysis is facilitated by the recent editions of the Stata program (Stata 2013) which has established routines for statistical analysis and graphical presentations of such data. In the recent years, the conduct of meta-analyses in the field of nutritional epidemiology has increased considerably compared to the past. This development is facilitated by the many nutritional epidemiological studies generated from the cohorts established around the beginning of this century. It is important to recognize that a meta-analysis is not necessarily based on a systematic review of all available literature, thus the possibility of bias may still arise.

It is nevertheless often not easy to summarize the evidence by the MAL approach. It was shown that the presentation of results from risk analysis for dietary exposure can be very different. Usually, the researchers present the data in such a way that the main message can be well captured by the readers under the usual space restrictions of journals. We have shown in the previous sections that many ways exist to describe the relation between a dietary exposure and outcome. Depending on the aim of a meta-analysis, the metric of interest has to be chosen. This will determine whether a study is included or not.

A nice example for summarizing the evidence by meta-analysis has been provided by the World Cancer Research Fund (WCRF) in its second report on Food, Nutrition, Physical Activity and the Prevention of Cancer and the subsequent systematic literature reviews (WCRF 2007). Particularly, in the material from the systematic literature reviews published along with the book, several metrics were used to describe the association: (1) by categories (comparing the lowest vs. the highest), (2) by using units of exposure, and (3) by displaying association along the range of exposure from the various studies.

Published data regarding dietary exposure are still liable to gross biases compared to the truth and thus even carefully conducted meta-analyses of nutritional epidemiological studies cannot be directly transformed into quantitative recommendations. Meta-analyses are providing hints regarding the direction of an effect but currently the data are not solid enough for immediate quantitative public health advice.

Pooled analyses have the advantage that the original data can be calibrated and uniform quality criteria can be applied to all studies included. A recent example of this is the analysis of trials of the effects of supplementation with multiple micronutrients compared with iron and folic acid supplementation during pregnancy (Margetts et al. 2009). Having access to the raw data makes it possible to apply the same inclusion and exclusion criteria and to adjust for gestational age in a comparable manner across all studies (Margetts et al. 2009). There are several initiatives to establish large databases with data from nearly all cohort studies, particularly those conducted in the USA and Europe, for pooled analyses. Depending on the degree of quality control and calibration efforts, such joint analyses could provide meaningful quantitative summary estimates in the future.

44.8 Nutritional Epidemiology in Public Health Practice

Here, we discuss the use of nutritional epidemiology methods for descriptive epidemiology and public health nutrition practice rather than research. In particular, we focus on national dietary surveys and the assessment of population intakes with respect to external references because there have been several important changes in methodological approach and terminology in recent years.

44.8.1 Nutrition Surveys: Assessing the Usual Intake of a Population

With the notable exception of the UK, countries that have conducted national surveys investigating dietary intake have generally collected 1 day of intake from participants, usually as a 24-h recall but occasionally as a 1-day record. In the UK, 4–7 days of records have been collected in all national diet and nutrition surveys (recent cycle began in 1986 with adults and has covered all age groups, up to repeating adults in 2002).

The external references (see below) are based on the assumption that usual intake has been assessed in the individuals participating in the survey. If national survey data collected a single day of information from each individual, it will be difficult to compare such data to a reference due to lack of information on the intraindividual variation and wrong conclusions about the prevalence of specific exposures might

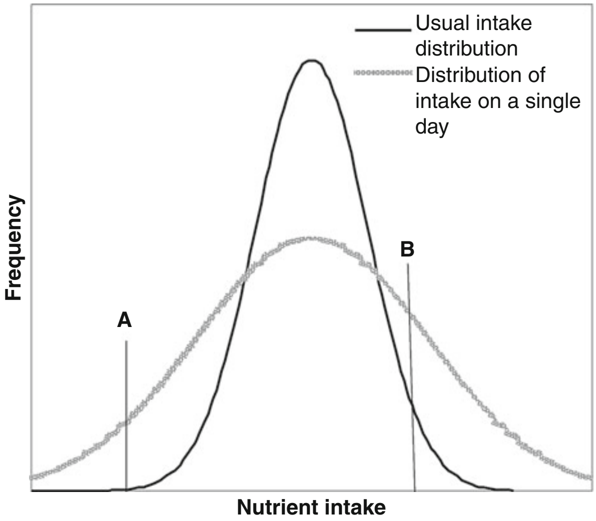


Fig. 44.3 The distribution of the usual intake of individuals has a narrower standard deviation than the intake on a single day. When cut-offs (e.g., *A* or *B*) are based on the assumption that usual intake has been measured, the incorrect prevalence will be obtained if single-day data is used

be drawn. If the aim is, for example, to describe the proportion of the population with low intakes, then using single-day data will overestimate the true prevalence when the population average lies above the cut-off (cut-off A in Fig. 44.3), but will underestimate the true prevalence if the population average lies below the cut-off (cut-off B in Fig. 44.3). Conversely, if the interest is in describing the proportion with high intakes, then single-day data will underestimate and overestimate the true prevalence, respectively. The extent of the error will depend on the location of the cut-off with respect to the population average and also the ratio of the standard deviations of the single day and usual intake distributions. This ratio is lower for those nutrients found in a wide range of foods which are eaten frequently (e.g., macronutrients such as protein) and highest for those nutrients which are found in a small range of foods where the alternatives have different nutrient contents (e.g., the vitamin A content of vegetables varies enormously).

There are two alternative ways of dealing with this problem. One is to take the strategy used in the UK and obtain multiple days of intakes from all participants in a survey. This increases the cost, logistical complexity, and respondent burden. The second way is to obtain an estimate of the ratio of the standard deviations of the two distributions and to correct the distribution obtained from all participants using the ratio (Sempos et al. 1991). As dietary intakes and patterns vary with age and sex, it would be important to do this work for all important subgroups in the population and not to obtain one ratio, for example, in adults that is then applied to all other groups. This method is obviously cheaper. Its main drawback is the fact that the usefulness of the survey data for other purposes is limited.

Table 44.3 Data from other studies showing the range of intakes across fourths or fifths for different ways of expressing nutrient intakes

Author	Fifths				
Nutrient and method	Lowest	2	3	4	5
Willett et al. (1987) – data from 4 one-week diet records in 179 women					
Total fat					
Mean absolute (gram)	47	59	68	75	98
Energy-adjusted (g)[a]	56	64	69	72	78
Mean % energy	32	36	39	41	44
Cholesterol					
Mean absolute (milligram)	204	262	325	345	436
Mean energy-adjusted (mg)[a]	216	268	301	337	423
Mean mg/1,000 kilocalories	136	166	188	212	276
Brisson et al. (1989) – data from a food frequency questionnaire					
Total fat					
Median absolute (g)	56	78	97	132	–
Median energy-adjusted (g)[a]	78	88	96	106	–
Median % energy	33	36	38	40	–
Cholesterol					
Median absolute (mg)	201	291	362	487	–
Median energy-adjusted (mg)[a]	249	312	357	442	–
Median mg/1,000 kcal	127	144	157	181	–
Fiber					
Median absolute (g)	4.3	6.5	8.2	11.1	–
Median energy-adjusted (g)[a]	4.8	6.7	8.2	10.4	–

[a] Each person's residual has been added to the mean for the population for that nutrient

One-way analysis of variance allowing for random effects and repeated measures can be used to separate the between-person variance (s_b) from the total variance (s_{total}) in the subsample with multiple measures (Mackerras 1998). Then the ratio can be used to adjust each individual's single-day value (x_i) in the main survey relative to the sample mean ($\overline{X}$) (Sempos et al. 1991):

$$\text{adjusted value} = \overline{X} + (x_i - \overline{X})(s_b/s_{total}).$$

Depending on the nutrient, the 10th to 90th centile range of the corrected distribution could be as little as 66% of the 10th to 90th centile range of the uncorrected distribution (cf. Sect. 44.3.3 and Table 44.3). It is important to note that although the corrected population distribution was created by applying a factor to each individual's data point, the resulting values for each individual cannot be interpreted as each individual's true usual intake (Guenther et al. 1997; Murphy 2003). They are only correct on average, and this is sufficient to generate the corrected population distribution. If the usual intake has to be estimated for each individual (for addressing other purposes of the survey), multiple days of intake data must be collected from each individual.

Just as nutrient intake varies from day to day, so does the intake of foods and non-nutrient food components. Given current recommendations to increase intake of foods such as fruit and vegetables, it is important for dietary surveys that collected single-day data are corrected using the same approach for each food. Although there are a number of studies describing the extent of the variability of nutrient intake (e.g., Beaton et al. 1979, 1983; Nelson et al. 1989), there are few reports supplying the same information for food intake (Palaniappan et al. 2003).

44.8.2 Measuring Nutritional Exposures in Specific Groups

The characteristics of the target population and the circumstances in which they live influence the approach that has to be taken in gathering information about nutritional exposures. An approach that works in one community, or sector of society, may not work in another one. People from different cultures have different ways of conceptualizing and expressing what is important for them. For example, people who gather their food from the wild will often have very detailed names for all the edible foods but often have only one name that describes all the other foods that they do not eat.

44.8.3 References for Assessing Dietary Intake in Populations

For many years, “recommended” intakes of nutrients have been promulgated by various national bodies. However, the correct use of these figures has never been very clear. This situation is probably related to the general problem that diagnostic criteria for assessing individuals are often misinterpreted as indicators for population surveillance, or vice versa. A typical use of national dietary surveys is to compare the results to external references such as “recommended intakes.”

It is important to realize that all committees agree that the “recommended” figures they produce are expressed as daily amounts for convenience only. All committees agree that it is long-term diet that is important and that the “recommended” figures should never be compared to a single day of intake for either populations or individuals. As shown in Fig. 44.3, the incorrect prevalence will be found if they are compared to single-day data. The following discussion assumes that usual intake distributions are available for the population.

Until about 10 years ago, most countries have set up only one type of dietary reference figure for each age-sex group for each nutrient. At present, this figure is called the Reference Nutrient Intake (RNI) in the UK, the Population Reference Intake in the European Union, the Recommended Dietary Allowance in the USA and Canada, and the Recommended Dietary Intake in Australia (Department of Health 1991; The Scientific Committee for Food 1993; Food and Nutrition Board 2000; Truswell et al. 1990). We will refer to all of them as the RNI. The RNI is generally set well above the average requirement so that, if everyone ate this amount, there would be a low probability of deficiency in the population. Sometimes, there was enough information to set the RNI at two standard deviations above the Estimated Average Requirement (EAR). Often, the committees just added a large safety margin. As the margin added varied by nutrient, it was not possible to express the RNI as a constant multiple of the underlying average requirement.

The UK was the first country to set multiple values when it revised its dietary references (Department of Health 1991) and specified the location of the EAR in addition to the RNI which is referring to the 97th centile value, not the average. The European Union then followed, naming its equivalent figures Average Requirements (The Scientific Committee for Food 1993) and more recently the USA and Canada have also described the EAR that their recommended dietary allowances were derived from (Food and Nutrition Board 2000).

Advising individuals to have intakes that meet the RNI is sensible because this advice carries a low probability of suggesting inadequate intakes for that person. However, could the RNI be used to assess the adequacy of population intakes? In earlier times when the EAR was not specified, two methods were used quite commonly. They were to determine whether the mean intake is equal to or above the RNI or to determine the proportion falling below the RNI. Neither of these approaches is satisfactory (Beaton 1999).

Because the standard deviation of the population intake distribution is generally much wider than the standard deviation of the requirement distribution (Beaton 1999), many people will appear to have intakes below their requirements when the median population intake lies on the RNI (RNI(a) in Scenario 1, Fig. 44.4). The extent to which this happens depends on how the RNIs were set up in the past. If they were set up by adding a safety margin that was larger than twice the $\text{SD}_{\text{requirement}}$ (SD = standard deviation), then the greater margin will lead to less prevalence of inadequate intake (RNI(b) in Scenario 1, Fig. 44.4). It is true that, if everyone has an intake above the RNI, the prevalence of inadequate intake will be essentially 0 (Scenario 2, Fig. 44.4). However, many people will be consuming much more than their personal requirements, and it is clear that there could be some overlap between the two distributions that is still consistent with a low probability of inadequate intake.

A third approach that has been used in the past is to calculate the proportion of the population with intakes below 70% of the RNI based on the assumption that the requirement distribution had a coefficient of variation of 20%. As this was rarely the case, this approach would overestimate the true prevalence of inadequate intake for some nutrients and underestimate the true prevalence for other nutrients. Consequently, this approach would not necessarily reveal which nutrient is in shortest supply in the population.

So the question is how much overlap there can be between the requirement and intake distributions before the prevalence of inadequate intakes is unacceptably high? A method for multiplying the two probability distributions was described in 1986 (Subcommittee on Criteria for Dietary Evaluation 1986), and the powerful computers of today can do the calculations easily. However, if the requirement distribution is symmetrical and the intake distribution is approximately normal, then a shortcut method can be used: if the intake mean is at least $\text{EAR}+2{*}\text{SD}_{\text{intake}}$

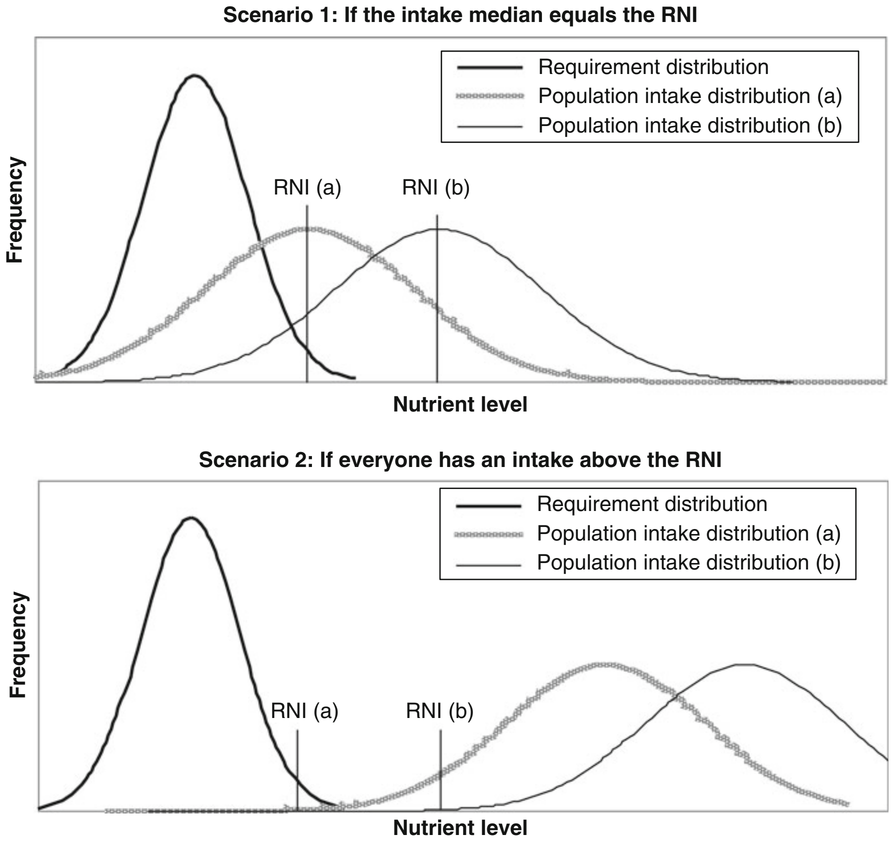


Fig. 44.4 Illustration of how the Reference Nutrient Intake (RNI) has been used in the past to assess population intakes, assuming two different RNIs that had been set in different ways (Adapted from Beaton 1999)

(with 2 being approximately the 97.5% quantile of a standard normal distribution) higher than the EAR, then the prevalence of inadequate intakes will be 2.5% or less (Fig. 44.5). In other words, if the assumptions are met, the proportion below the EAR is the proportion with inadequate intakes, although no definite statements can be made about the individuals with intakes below the EAR. Other multiples of the SD_{intake} could be used if other criteria are desired. Note that the standard deviation in this formula is for the intake distribution, not the requirement distribution, and the reference point is the EAR and not the RNI. The details of this approach and further information on other situations, such as when the requirement distribution is asymmetrical, can be found elsewhere (Subcommittee on Criteria for Dietary Evaluation 1986; Beaton 1999; Food and Nutrition Board 2000).

As a final note, Fig. 44.5 has assumed that intake distributions are normal. If these must be transformed, then care needs to be taken to apply the correction factors appropriately. The above comments about the RNI do not apply to energy, because

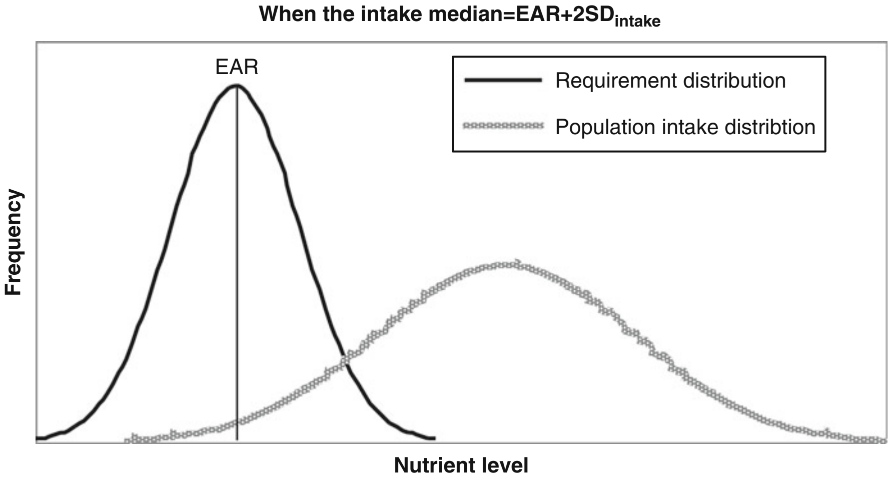


Fig. 44.5 If the mean population intake is 2*SD$_{intake}$ (*SD* = standard deviation) above the Estimated Average Requirement (EAR), then the prevalence of inadequate intakes in the population is 2.5%

the energy recommendations are set at the EAR and not at two SD above the EAR. It is also important to examine how words are used in each country, because the same word may be used to signify different concepts, and to realize that definitions may even vary within a document. Values for the EAR and RNI may vary between countries, either because they have been based on different indicators of nutritional status or on a different selection of literature sources or because they are defined as mg/kg while the mean body weight varies between populations.

44.8.4 Impact of Underreporting of Intake

Because people tend to underreport their intake by most dietary methods (Black et al. 1991), the data collected in a national survey will tend to underestimate total intake. For example, when the criteria of Goldberg et al. (1991) for assessing a 24-h intake were applied to the 1995 Australian survey, 11.9% of men and 20.6% of adult women had energy intakes that were implausible. Compared to the total population surveyed, those with plausible intakes had higher intakes of all nutrients (Table 44.4). Clearly leaving those with implausible intakes in the data can affect the results of the analysis, especially for some of the minerals. Therefore analyses of national surveys that have not excluded these individuals would tend to overestimate the prevalence of low intakes of most nutrients. However, one problem with simply excluding them is that people may underreport some foods differently from others. If energy-containing foods are underreported to a greater extent than low-energy foods such as fruit and vegetables, then excluding the implausible reporters may incorrectly inflate intakes of nutrients such as vitamins A and C and to some extent

Table 44.4 Comparison of median nutrient intakes in the total survey sample and persons with plausible EI/BMR ratios, adults 19 years and older, by sex. 1995 National Nutrition Survey, Australia (ABS & HEALTH 1998)

	Men		Women	
Nutrient	Total	Plausible[a]	Total	Plausible[a]
Energy (kilojoule)	10, 377	10, 997	7, 083	7, 824
Protein (g)	100.1	106.4	69.5	76.3
Total fat (g)	89.8	96.4	61.6	70.3
Dietary fiber (g)	23.8	25.1	18.9	20.5
Vitamin A (ug retinol equivalents)	941	1, 012.9	753.6	833.0
Thiamin (mg)	1.7	1.8	1.2	1.3
Riboflavin (mg)	2.0	2.2	1.6	1.7
Niacin equivalents (mg)	47.1	49.8	32.3	35.1
Folate (mg)	285.3	299.6	216.7	232.6
Vitamin C (mg)	102.9	110.2	85.4	92.2
Calcium (mg)	827.3	891.4	663.1	737.6
Magnesium (mg)	360.3	380.8	266.9	291.1
Iron (mg)	15.2	16.1	11.1	12.2
Zinc (mg)	12.8	13.6	8.7	9.6

[a]Plausible intakes: 24-h intakes with an energy intake/basal metabolic rate ratio of 0.9 or greater are above the lower bound of the 95% confidence interval in a weight-stable individual undertaking light activity (Goldberg et al. 1991)

folate and fiber. A further problem is that equivalent criteria for defining those with implausibly high reports are not available.

44.8.5 Consequences of Within-Person Variability in Other Areas of Public Health Nutrition Practice

As noted above, intraindividual variation in dietary measures is not due to the subjective nature of the measure. The error related to reporting or measurement occurs in addition to the underlying instability of the parameter being measured. Many “objective” measures also vary every time they are measured. Blood pressure is a notable example – it can vary from minute to minute. Even biochemical measures also exhibit within-person variability. Some parameters have a regular variation (e.g., diurnal variation) in addition to random variation. The particular problems with dietary measures occur because the within-person variance is larger, generally much larger, than the between-person variance. However, the observed prevalence can be affected even for parameters where the within-person variance is smaller than the between-person variance. Looker et al. (1990) found that the prevalence of impaired iron status (based on mean corpuscular volume, transferrin saturation, and erythrocyte protoporphyrin) was 10% in a national survey using the single measure data. This was reduced to 4% when corrected for within-person variation.

The within-person variation is why many parameters are measured more than once before a diagnosis and treatment decision can be made in the clinical setting. Irwig et al. (1991) showed that the interpretation of a single measure of cholesterol level in a client depends on knowing the underlying population distribution and that single measures cannot be used to assess whether the client's true underlying average has changed since the previous measurement. These observations can be generalized to other characteristics as well.

Program evaluation is another area where taking a single measure may yield misleading results because of within-person variability and measurement error. People may be eligible to enter a program because their level of a characteristic is below a cut-off. In this case, the mean level of the characteristic in the eligible group will be higher when it is measured again than it was at baseline even if the program has no effect. Similarly, if people are selected because their characteristic is above a cut-off, then the group mean will be lower on the second occasion, even if the program is ineffective. In both cases, the mean value of the second measurement is closer to the mean in the total unselected population than the mean of the first measurement. This effect occurs because of a statistical phenomenon called regression to the mean (Davis 1976; Newell and Simpson 1990; Bland and Altman 1994). If the evaluation uses a randomized control design, then the effects of regression to the mean and other explanations such as seasonal or secular trends occur in both the control and the intervention group. The artefact can thus be avoided by comparing the follow-up level in the intervention group to the follow-up level in the control group instead of comparing the follow-up to the baseline levels in the intervention group only. Because randomized studies are perceived as very complex, it may be tempting to conduct a non-randomized study and using ineligible people (i.e., those whose values did not meet the criterion for eligibility) as a control group, but this will not allow the regression to the mean effect to be detected. Sometimes, a specific subgroup is not selected at the outset, but the total group may be divided up in the analyses and authors may report that those with the most extreme values at baseline benefited most from the intervention. If there is no randomized control arm, this sort of finding is questionable because the results may be simply due to regression to the mean rather than to the intervention (Vickers and Altman 2001).

Whether a national survey provides a useful source of information for studying diet-disease relationships depends on exactly how it is carried out. Even though it may be possible to demonstrate associations, national surveys which assess dietary intake and outcome markers such as blood pressure and hemoglobin levels are cross-sectional surveys, even if the information is collected over a few days. This inevitably limits causal inference. It is not possible to determine the temporal relationship between, e.g., concentration of blood lipids or blood pressure and consumption of fat type or sodium. Sometimes, a population involved in a national survey may be followed up beyond the survey. Even though this converts a cross-sectional survey into a cohort study, the extent of the analysis would depend on the initial measurements obtained. If only a single 24-h recall was obtained, then correlation and regression coefficients will be attenuated (Liu et al. 1978; Sempos et al. 1985) as previously described.

44.9 Conclusions

There are few health outcomes for which nutrition does not play a direct or indirect role in causation and therefore in disease prevention. Worldwide, there is a complex mix of problems of over- and undernutrition, often stratified by education or economic groups (World Health Organisation 2011). Some of these nutrition-related problems are clear and simply require the political will to be dealt with. Others are more complex and the correct way to solve the problem may not be obvious. This is where nutritional epidemiology has to play a critical role. In order to improve and maintain public health, it is important to provide a strong evidence base to guide action. This is particularly the case where there are vested interests that may not want the dietary patterns to change. In order to justify and support such changes, it is essential that the scientific evidence supports policy. The methods of nutritional epidemiology guide that evidence base.

The major concern in nutritional epidemiology is how to define and measure with required accuracy the relevant measure of exposure, free from bias. Because diet and other behaviors are complex and interrelated, it is important, in both the design and interpretation of studies, to understand how this complexity may affect the results of the study. These issues cannot simply be resolved by statistical adjustment, it is essential to have an understanding of the underlying social, psychological, and biological factors.

Because most studies are of limited statistical power, there is a growing tendency to undertake meta-analyses using pooled data. Before data are pooled, it is important to assess whether there are methodological differences between studies. This is more than assessing the heterogeneity using a statistical technique. It is important to investigate whether the differences in dietary assessment methods and ways of presenting data allow such pooling. The range of exposure in the referent category and the way data are subdivided into categories will all affect the prevalence of the exposure and the size of the risk estimate in each study.

Acknowledgments We would like to thank Dorothy Mackerras for her input to the previous chapter on Nutritional Epidemiology of this book. Some sections of this revised text are largely unchanged from the first edition.

References

ABS & HEALTH (1998) National Nutrition Survey. Nutrient intakes and physical measurements Australia 1995. Catalogue No 4805.0. Australian Bureau of Statistics, Canberra

Amanatidis S, Mackerras D, Simpson JM (2001) Comparison of two frequency questionnaires for quantifying fruit and vegetable intake. Public Health Nutr 4:233–239

Arab L, Tseng CH, Ang A, Jardack P (2011) Validity of a multipass, web-based, 24-hour self-administered recall for assessment of total energy intake in blacks and whites. Am J Epidemiol 11:1256–1265

Bälter KA, Bälter O, Fondell E, Lagerros YT (2005) Web-based and mailed questionnaires: a comparison of response rates and compliance. Epidemiology 4:577–579

Bates (1997a) Bioavailability of riboflavin. Eur J Clin Nutr 51(Suppl 1):S38–S42

Bates (1997b) Bioavailability of vitamin C. Eur J Clin Nutr 51(Suppl 1):S28–S33

Beaton GH (1999) Recommended dietary intakes: individuals and populations. In: Shils ME, Olson JA, Shike M, Ross AC (eds) Modern nutrition in health and disease, 9th edn. Williams and Wilkins, Baltimore

Beaton GH, Milner J, Corey P, McGuire V, Cousins M, Stewart E, de Ramos M, Hewitt D, Grambsch PV, Kassim N, Little JA (1979) Sources of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation. Am J Clin Nutr 32:2546–2549

Beaton GH, Milner J, McGuire V, Feather TE, Little JA (1983) Source of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation. Carbohydrate sources, vitamins, and minerals. Am J Clin Nutr 37:986–995

Bell S, Pakkala H, Finglas MP (2012) Towards a European food composition data interchange platform. Int J Vitam Nutr Res 82:209–215

Binder H, Sauerbrei W, Royston P (2013) Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. Stat Med 32:2262–2277

Bingham SA, Luben R, Welch A, Wareham N, Khaw KT, Day N (2003) Are imprecise methods obscuring a relation between fat and breast cancer? Lancet 362:212–214

Biobank UK (2013) Improving the health of future generations. Questions on diet. http://www.ukbiobank.ac.uk/wp-content/uploads/2011/07/diet_questionnaire.pdf. Accessed 14 May 2013

Black AE, Goldberg GR, Jebb SA, Livingstone MB, Cole TJ, Prentice AM (1991) Critical evaluation of energy intake data using fundamental principles of energy physiology: 2. Evaluating the results of published surveys. Eur J Clin Nutr 45:583–599

Bland JM, Altman DG (1994) Some examples of regression towards the mean. BMJ 309:780

Börnhorst C, Huybrechts I, Ahrens W, Eiben G, Michels N, Pala V, Molnár D, Russo P, Barba G, Bel-Serrat S, Moreno LA, Papoutsou S, Veidebaum T, Loit HM, Lissner L, Pigeot I (2013) Prevalence and determinants of misreporting among European children in proxy-reported 24 h dietary recalls. Br J Nutr 109:1257–1265

Brisson J, Verreault R, Morrison AS, Tennina S, Meyer F (1989) Diet, mammographic features of breast tissue, and breast cancer risk. Am J Epidemiol 130:14–24

Brustad M, Skeie G, Braaten T, Slimani N, Lund E (2003) Comparison of telephone vs. face-to-face interviews in the assessment of dietary intake by the 24 h recall EPIC SOFT program – the Norwegian calibration study. Eur J Clin Nutr 57:107–113

Burke B (1947) The dietary history as a tool in research. J Am Diet Assoc 23:1041–1046

Carroll RJ, Midthune D, Subar AF, Shumakovich M, Freedman LS, Thompson FE, Kipnis V (2012) Taking advantage of the strengths of 2 different dietary assessment instruments to improve intake estimates for nutritional epidemiology. Am J Epidemiol 175:340–347

Convey JM, Ingwersen LA, Moshfegh AJ (2004) Accuracy of dietary recall using the USDA five-step multiple-pass method in men: an observational validation study. J Am Diet Assoc 4:595–603

Daniel CR, Cross AJ, Graubard BI, Hollenbeck AR, Park Y, Sinha R (2011) Prospective investigation of poultry and fish intake in relation to cancer risk. Cancer Prev Res (Phila) 11:1903–1911

Daugherty BL, Schap TE, Ettienne-Gittens R, Zhu FM, Bosch M, Delp EJ, Ebert DS, Kerr DA, Boushey CJ (2012) Novel technologies for assessing dietary intake: evaluating the usability of a mobile telephone food record among adults and adolescents. Med Internet Res 2:e58

Daurès JP, Gerber M, Scali J, Astre C, Bonifacj C, Kaaks R (2000) Validation of a food-frequency questionnaire using multiple-day records and biochemical markers: application of the triads method. J Epidemiol Biostat 5:109–115

Davis CE (1976) The effect of regression to the mean in epidemiologic and clinical studies. Am J Epidemiol 104:493–498

Department of Health (1991) Report of Health and Social Subjects 41. Dietary reference values for food energy and nutrients for the United Kingdom. Report of the Panel on Dietary Reference Values of the Committee on Medical Aspects of Food Policy. HMSO, London

FAO (2013) FAOSTAT. http://faostat.fao.org/site/354/default.aspx. Accessed 28 Mar 2013

Food and Nutrition Board, Institute of Medicine (2000) Dietary Reference Intakes: applications in dietary assessment. National Academy Press, Washington, DC. http://www.nap.edu/catalog/9956.html. Accessed 29 Apr 2004

Freedman LS, Midthune D, Carroll RJ, Tasevska N, Schatzkin A, Mares J, Tinker L, Potischman N, Kipnis V (2011) Using regression calibration equations that combine self-reported intake and biomarker measures to obtain unbiased estimates and more powerful tests of dietary associations. Am J Epidemiol 174:1238–1245

García-Segovia P, González-Carrascosa R, Martínez-Monzó J, Ngo J, Serra-Majem L (2011) New technologies applied to food frequency questionnaires: a current perspective. Nutr Hosp 26:803–806

George SM, Thompson FE, Midthune D, Subar AF, Berrigan D, Schatzkin A, Potischman N (2012) Strength of the relationships between three self-reported dietary intake instruments and serum carotenoids: the Observing Energy and Protein Nutrition (OPEN) Study. Public Health Nutr 15:1000–1007

German Institute of Human Nutrition (2013) Questionnaire and forms. https://sms.dife.de/tools/current/de. Accessed 14 May 2013

Goldberg GR, Black AE, Jebb SA, Cole TJ, Murgatroyd PR, Coward WA, Prentice AM (1991) Critical evaluation of energy intake data using fundamental principles of energy physiology: 1. Derivation of cut-off limits to identify under-recording. Eur J Clin Nutr 45:569–581

Goran MI, Astrup A (2002) Energy metabolism. In: Gibney MJ, Vorster HH, Kok FJ (eds) Introduction to human nutrition. Blackwell Publishing, Oxford

Guenther PM, Kott PS, Carriquiry AL (1997) Development of an approach for estimating usual nutrient intake distributions at the population level. J Nutr 127:1106–1112

Haraldsdóttir J, Tjønneland A, Overvad K (1994) Validity of individual portion size estimates in a food frequency questionnaire. Int J Epidemiol 23:786–796

Hosmer DW, Lemeshow S (1989) Applied logistic regression. Wiley, New York

Hu FB (2002) Dietary pattern analysis: a new direction in nutritional epidemiology. Curr Opin Lipidol 13:3–9

Illner AK, Nöthlings U, Wagner K, Ward H, Boeing H (2010) The assessment of individual usual food intake in large-scale prospective studies. Ann Nutr Metab 2:99–105

Illner AK, Harttig U, Tognon G, Palli D, Salvini S, Bower E, Amiano P, Kassik T, Metspalu A, Engeset D, Lund E, Ward H, Slimani N, Bergmann M, Wagner K, Boeing H (2011) Feasibility of innovative dietary assessment in epidemiological studies using the approach of combining different assessment instruments. Publ Health Nutr 14:1055–1063

Illner AK, Freisling H, Boeing H, Huybrechts I, Crispim SP, Slimani N (2012) Review and evaluation of innovative technologies for measuring diet in nutritional epidemiology. Int J Epidemiol 4:1187–1203

Irwig L, Glasziou P, Wilson A, Macaskill P (1991) Estimating an individual's true cholesterol level and response to intervention. JAMA 266:1678–1685

Kipnis V, Freedman LS, Brown CC, Hartman AM, Schatzkin A, Wacholder S (1993) Interpretation of energy adjustment methods for nutritional epidemiology. Am J Epidemiol 137:1376–1380

Koebnick C, Wagner K, Thielecke F, Dieter G, Hohne A, Franke A, Garcia AL, Meyer H, Hoffmann I, Leitzmann P, Trippo U, Zunft HJ (2005) An easy-to-use semiquantitative food record validated for energy intake by using doubly labelled water technique. Eur J Clin Nutr 59:989–995

Kohlmeier L (1995) Future of dietary exposure assessment. Am J Clin Nutr 61:702S–709S

Kolonel LN, Henderson BE, Hankin JH, Nomura AMY, Wilkens LR, Pike MC, Stram DO, Monroe KR, Earle ME, Nagamine FS (2000) A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. Am J Epidemiol 151:346–357

l'Étude Nutrinet-Santé (2013) https://www.etude-nutrinet-sante.fr. Accessed 28 Mar 2013

Liu K, Stamler J, Dyer A, McKeever J, McKeever P (1978) Statistical methods to assess and minimize the role of intra-individual variability in obscuring the relationship between dietary lipids and serum cholesterol. J Chronic Dis 31:399–418

Liu B, Young H, Crowe FL, Benson VS, Spencer EA, Key TJ, Appleby PN, Beral V (2011) Development and evaluation of the Oxford WebQ, a low-cost, web-based method for assessment of previous 24 h dietary intakes in large-scale prospective studies. Public Health 11:1998–2005

Looker AC, Sempos CT, Liu KA, Johnson CL, Gunter EW (1900) Within-person variance in biochemical indicators of iron status: effects on prevalence estimates. Am J Clin Nutr 52: 541–547

Mackerras D (1998) Within- and between-subject variability. In: Kerr CB, Taylor R, Heard G (eds) Handbook of public health methods. McGraw-Hill, Sydney

Margetts BM, Nelson M (eds) (1997) Design concepts in nutritional epidemiology, 2nd edn. Oxford University Press, Oxford

Margetts BM, Vorster HH, Venter CS (2003) Evidence based nutrition: the impact of information and selection bias on the interpretation of individual studies. SAJCN 16:79–87

Margetts BM, Fall CH, Ronsmans C, Allen LH, Fisher DJ, Maternal Micronutrient Supplementation Study Group (2009) Multiple micronutrient supplementation during pregnancy in low-income countries: review of methods and characteristics of studies included in the meta-analyses. Food Nutr Bull 30(4 Suppl):S517–S526

Mark SD, Thomas DG, Decarli (1996) A measurement of exposure to nutrients: an approach to the selection of informative. Am J Epidemiol 5:514–521

Medical Research Council (MRC) (2013) Diet and physical activity measurement toolkit. http://dapa-toolkit.mrc.ac.uk/. Accessed 14 May 2013

Murakami K, Sasaki S, Takahashi Y, Uenishi K, Watanabe T, Kohri T, Yamasaki M, Watanabe R, Baba K, Shibata K, Takahashi T, Hayabuchi H, Ohki K, Suzuki J (2008) Sensitivity and specificity of published strategies using urinary creatinine to identify incomplete 24-h urine collection. Nutrition 24:16–22

Murphy SP (2003) Collection and analysis of intake data from the integrated survey. J Nutr 133:585S–589S

Naska A, Oikonomou E, Trichopoulou A, Wagner K, Gedrich K (2007) Estimations of daily energy and nutrient availability based on nationally representative household budget survey data. The Data Food Networking (DAFNE) project. Public Health Nutr 12:1422–1429

National Cancer Institute (2013) Risk factor monitoring and methods. http://riskfactor.cancer.gov/diet/usualintakes/method.html. Accessed 28 Mar 2013

Nelson M, Black AE, Morris JA, Cole TJ (1989) Between- and within-subject variation in nutrient intake from infancy to old age: estimating the number of days required to rank dietary intakes with desired precision. Am J Clin Nutr 50:155–167

Newell D, Simpson J (1990) Regression to the mean. Med J Aust 153:166–168

Ngo J, Engelen A, Molag M, Roesle J, García-Segovia P, Serra-Majem L (2009) A review of the use of information and communication technologies for dietary assessment. Br J Nutr Suppl 2:S102–S112

Noethlings U, Hoffmann K, Bergmann MM, Boeing H (2003) European investigation into cancer and nutrition portion size adds limited information on variance in food intake of participants in the EPIC-Potsdam study. J Nutr 2:510–515

Nöthlings U, Hoffmann K, Bergmann MM, Boeing H (2007) Fitting portion sizes in a self-administered food frequency questionnaire. J Nutr 12:2781–2786

Nutrition Analysis Tool (NAT) (2013) NAT tools for good health. http://www.myfoodrecord.com Accessed 14 May 2013

Ovaskainen ML, Paturi M, Reinivuo H, Hannila ML, Sinkko H, Lehtisalo J, Pynnönen-Polari O, Männistö S (2008) Accuracy in the estimation of food servings against the portions in food photographs. Eur J Clin Nutr 62:674–681

Palaniappan U, Cue RI, Payette H, Gray-Donald K (2003) Implications of day-to-day variability on measurements of usual food and nutrient intakes. J Nutr 133:232–235

Poslusna K, Ruprich J, de Vries JH, Jakubikova M, van't Veer P (2009) Misreporting of energy and micronutrient intake estimated by food records and 24 hour recalls, control and adjustment methods in practice. Br J Nutr Suppl 2:S73–S85

Rosner B, Michels KB, Chen YH, Day NE (2008) Measurement error correction for nutritional exposures with correlated measurement error: use of the method of triads in a longitudinal setting. Stat Med 27:3466–3489

Royston P, Sauerbrei W (2008) Multivariable model-building: a pragmatic approach to Regression analysis based on fractional polynomials for modeling continuous variables. Wiley, Chichester

Sempos CT, Johnson NE, Smith EL, Gilligan C (1985) Effects of intraindividual and interindividual variation in repeated dietary records. Am J Epidemiol 121:120–130

Sempos CT, Looker AC, Johnson CL, Woteki CE (1991) The importance of within-person variability in estimating prevalence. In: Macdonald I (ed) Monitoring dietary intakes. Springer, Berlin

Sinha R (2002) An epidemiologic approach to studying heterocyclic amines. Mutat Res 506–507:197–204

Slimani N, Ferrari P, Ocké M, Welch A, Boeing H, Liere M, Pala V, Amiano P, Lagiou A, Mattisson I, Stripp C, Engeset D, Charrondière R, Buzzard M, Staveren W, Riboli E (2000) Standardization of the 24-hour diet recall calibration method used in the European prospective investigation into cancer and nutrition (EPIC): general concepts and preliminary results. Eur J Clin Nutr 12:900–917

Spiegelman D, McDermott A, Rosner B (1997) Regression calibration method for correcting measurement-error bias in nutritional epidemiology. Am J Clin Nutr 65:1179S–1186S

Stata (2013) Data analysis and statistical software. http://www.stata.com. Accessed 14 May 2013

Subar AF, Thompson FE, Kipnis V, Midthune D, Hurwitz P, McNutt S, McIntosh A, Rosenfeld S (2001) Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: the Eating at America's Table Study. Am J Epidemiol 12:1089–1099

Subar AF, Kipnis V, Troiano RP, Midthune D, Schoeller DA, Bingham S, Sharbaugh CO, Trabulsi J, Runswick S, Ballard-Barbash R, Sunshine J, Schatzkin A (2003) Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: the OPEN study. Am J Epidemiol 158:1–13

Subar AF, Dodd KW, Guenther PM, Kipnis V, Midthune D, McDowell M, Tooze JA, Freedman LS, Krebs-Smith SM (2006) The food propensity questionnaire: concept, development, and validation for use as a covariate in a model to estimate usual food intake. J Am Diet Assoc 10:1556–1563

Subar AF, Kirkpatrick SI, Mittl B, Zimmerman TP, Thompson FE, Bingley C, Willis G, Islam NG, Baranowski T, McNutt S, Potischman N (2012) The Automated Self-Administered 24-hour dietary recall (ASA24): a resource for researchers, clinicians, and educators from the National Cancer Institute. J Acad Nutr Diet 8:1134–1137

Subcommittee on Criteria for Dietary Evaluation (1986) Nutrient adequacy: assessment using food consumption surveys. National Academy Press, Washington, DC. http://www.nap.edu/books/0309036348/html. Accessed 28 Mar 2013

The Scientific Committee for Food (1993) Nutrient and energy intakes for the European Community. Thirty-first series: Food – science and techniques series. Office for Official Publications of the European Communities, Luxembourg. PDF available at http://ec.europa.eu/food/fs/sc/scf/reports_en.html. Accessed 21 Apr 2013

Thompson FE, Subar AF, Brown CC, Smith AF, Sharbaugh CO, Jobe JB, Mittl B, Gibson JT, Ziegler RG (2002) Cognitive research enhances accuracy of food frequency questionnaire reports: results of an experimental validation study. J Am Diet Assoc 102:212–225

Thompson FE, Midthune D, Subar AF, Kahle LL, Schatzkin A, Kipnis V (2004) Performance of a short tool to assess dietary intakes of fruits and vegetables, percentage energy from fat and fibre. Public Health Nutr 7:1097–1105

Tooze JA, Krebs-Smith SM, Troiano RP, Subar AF (2012) The accuracy of the Goldberg method for classifying misreporters of energy intake on a food frequency questionnaire and 24-h recalls: comparison with doubly labeled water. Eur J Clin Nutr 66:569–576

Tran KM, Johnson RK, Soultanakis RP, Matthews DE (2000) In-person vs. telephone-administered multiple-pass 24-hour recalls in women: validation with doubly labeled water. J Am Diet Assoc 100:777–783

Truswell AS, Dreosti IE, English RM, Palmer N, Rutishauser IHE (eds) (1990) Recommended nutrient intakes. Australian papers. Australian Professional Publications, Mosman

van Staveren WA, Burema J, Livingstone MB, van den Broek T (1996) Evaluation of the dietary history method used in the SENECA Study. Eur J Clin Nutr Suppl 2:S47–S55

Vereecken CA, Covents M, Sichert-Hellert W, Alvira JM, Le Donne C, De Henauw S, De Vriendt T, Phillipp MK, Béghin L, Manios Y, Hallström L, Poortvliet E, Matthys C, Plada M, Nagy E, Moreno LA, HELENA Study Group (2008) Development and evaluation of a self-administered computerized 24-h dietary recall method for adolescents in Europe. Int J Obes (Lond) 32(Suppl 5):S26–S34

Vickers AJ, Altman DG (2001) Analysing controlled trials with baseline and follow-up measurements. BMJ 323:1123–1124

Walker AM, Blettner M (1985) Comparing imperfect measures of exposure. Am J Epidemiol 121:783–790

Willett WC (1998) Nutritional epidemiology, 2nd edn. Oxford University Press, New York

Willett WC (2013) Nutritional epidemiology, 3rd edn. Oxford University Press, New York

Willett WC, Stampfer MJ (1986) Total energy intake: implications for epidemiologic analyses. Am J Epidemiol 124:17–27

Willett WC, Stampfer MJ, Colditz GA, Rosner BA, Hennekens CH, Speizer FE (1987) Dietary fat and the risk of breast cancer. N Engl J Med 316:22–28

World Cancer Research Fund/American Institute for Cancer Research (2007) Food, nutrition, physical activity, and the prevention of cancer: aglobal perspective. AICR, Washington, DC

World Health Organisation (2011) Global status report on non-communicable diseases. WHO, Geneva

Zhang S, Midthune D, Guenther PM, Krebs-Smith SM, Kipnis V, Dodd KW, Buckman DW, Tooze JA, Freedman L, Carroll RJ (2011) A new multivariate measurement error model with zero-inflated dietary data, and its application to dietary assessment. Ann Appl Stat 5:1456–1487

Reproductive Epidemiology

45

Jørn Olsen and Olga Basso

Contents

J. Olsen (✉)
Department of Public Health, Section for Epidemiology, Aarhus University, Aarhus C, Denmark

Department of Epidemiology, Center for Health Sciences (CHS), UCLA School of Public Health, Los Angeles, CA, USA

O. Basso
Department of Obstetrics and Gynecology and Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC, Canada

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_27,

45.1 Introduction

45.1.1 Reproductive Epidemiology: Reading Instructions

In writing this chapter, we assumed that the reader is familiar with the basic concepts in epidemiology. Therefore, you will not find any overview of different designs, measure of disease occurrence, standardizations, other ways of adjusting for confounders, or any general discussion on bias, confounding or on measuring effects. If you are not familiar with these topics, you should start by reading other parts of the book or turn to one of the many fine available textbooks.

Our intent is to point out the problems and methods that are of particular interest in reproductive epidemiology, with emphasis on selected topics in this area of research.

We focus upon methodology, and, for the most part, we avoid reporting on any "state-of-the-art" overview on what is known about specific exposures or endpoints. Such reviews are soon outdated and, furthermore, space does not permit them here.

By highlighting some of the aspects that set reproductive epidemiology somewhat apart from other areas of epidemiology, we hope to alert readers to the problems and options that this field presents and to illustrate how important it is to develop a highly critical outlook. Research in reproductive epidemiology, like research in general, is not to prove or confirm anything but rather to question, make critical appraises, and, of course, identify potential causal or preventive factors of reproductive diseases.

We use references to illustrate specific problems of methodological interest. We use more of our own work than could ever be justified on the basis of our modest contribution to the field. Our justification is that these references reflect our source of information for learning about the many problems inherent to reproductive epidemiology. We cannot rule out that they also reflect our inflated egos. Most of the references we have selected present information of methodological relevance. We also provide references to full textbooks in reproductive epidemiology.

We do not in any way claim to provide a complete overview of problems that you should be aware of as a student of reproductive epidemiology. We are not aware of all these problems ourselves: some are yet to be described and some have not yet caught our attention. Even describing all the specific problems that we are aware of would require more space than you, the reader, would like us to have. What we have tried to do is to describe the most important problems as we see them. Our choice is a subjective one, and it reflects our experience and interests.

Reproductive health was defined by the WHO at the Cairo conference in 1994 as

> ... a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity, in all matters relating to the reproductive system and its functions and processes.
>
> Reproductive Health, therefore, implies that people are able to have a satisfying and safe sex life and that they have the capability to reproduce and the freedom to decide if, when, and how often to do so. It also includes sexual health, the purpose of which is the enhancement of life and personal relationships, and not merely counseling and care related to reproductive and sexually transmitted diseases.

We will not cover all the possible topics related to this definition, nor do we intend to provide research methods for studying well-being or even happiness. We will limit our focus to the more traditional domain of epidemiology, namely, to the studies of determinants of diseases directly related to reproduction, as long as these studies can be applied to human populations. Although many diseases will have an effect on procreation through biological, psychological, or social mechanisms, we will restrict ourselves to studies dealing with fecundity, pregnancy, birth, and early markers of child health. Many of the diseases we describe only manifest themselves in the time period of reproduction, such as subfecundity, pre-eclampsia, and, of course, birth outcomes.

We are aware of the hypothesis that many adult chronic diseases may originate in the pre- and perinatal time period. Although we consider this to be one of the most exciting new areas of epidemiology, this aspect will not be dealt with here but in the chapter on ▶Life Course Epidemiology of this handbook.

Our experience mainly stems from research in Europe and the USA, and we do not cover research problems of particular relevance to developing countries. Since reproductive health problems are usually larger in these countries, we are aware that this is a major shortcoming, and our only excuse is our limited experience in this field. We do want to stress, however, that research in developing parts of the world should also be based upon sound methods as the ones we advocate in this chapter. We do not believe in low-quality research anywhere, but we accept that circumstances may set limitations for what can be done.

Many diseases of the reproductive organs, such as cancer or infections, may have an effect on reproduction if the diseases appear before or during reproductive age. In most cases, studying the determinants of these diseases will be similar to studying determinants of other diseases, and, as such, they are not pertinent to the analysis in this chapter.

Our aim is not to provide a cookbook for research in reproductive health but to make readers aware of some of the aspects they should pay attention to. Associations come in many shapes, and they are not always what they pretend to be. “Be careful” is the main (and perhaps the only) message we want to convey, and if you are satisfied with that, you could stop here – if you want to know why you should be careful, please go on reading. We expect you to disagree on several occasions. Remember, you may be right and we may be wrong.

45.1.2 Reproductive Health: Specific Epidemiological Research Problems

Unlike epidemiologists studying cancer and chronic diseases, reproductive epidemiologists deal with an area that has been strongly shaped by evolution. Susceptibility to chronic diseases that do not affect the ability to reproduce would not be selected out in the course of evolution with the same pressure as would susceptibility to conditions that interfere with the ability to reproduce. Selective forces operate before and during pregnancy, even in the highly medicalized industrialized world; keeping this feature in mind when dealing with reproductive epidemiology is important when trying to understand events that occur in pregnancy.

Most epidemiologists deal with identifying determinants of diseases that may operate over time spans of varying length. The induction period may be very short (e.g., is a migraine attack triggered by a given nutritional component?) or may span many years (e.g., does prenatal exposure to smoking affect adult semen quality?) (Ramlau-Hansen et al. 2007). In reproductive health, we study determinants of factors that play a role for successful reproduction. Sometimes this involves studies that span generations, and many studies address couples (or families) rather than individuals. Unlike other fields of epidemiology, the individuals studied in reproductive epidemiology are often not ill in the common meaning of the word. Women who have repeated miscarriages or give birth to dead or severely malformed children are usually healthy themselves, and so are their partners. Women who become severely ill during pregnancy, for instance, with pre-eclampsia, often fully recover after the pregnancy ends, but they have a relatively high risk of experiencing pre-eclampsia again in their next pregnancy (and a somewhat increased risk of developing cardiovascular disease later in life). In general, risk of recurrence is relatively high for most reproductive outcomes. Many reproductive failures have a tendency to repeat themselves, suggesting that at least some of the causes are relatively stable in time.

In many cases, reproductive epidemiology deals with hidden phenomena that may represent serious disorders, which, however, may never come to light if a woman does not become pregnant. Furthermore, many women experience more than one pregnancy in their lifetime so that a pregnancy, rather than an individual, will often be the unit of analysis. All these elements, intrinsic to reproduction, provide interesting design options and challenging methodological problems, and they set reproductive epidemiology somewhat apart from other areas of epidemiology.

In evolutionary terms, successful reproduction means that the parents' genes are transmitted to viable offspring who will, eventually, produce offspring of their own. Part of this process is subject to epidemiological observation. Although the maternal and paternal evolutionary interests are similar as far as gene transmission is concerned, the parents' contributions – and stakes – are different. The mother provides the environment in which the fetus develops and is nourished during pregnancy and, often, also for a period of time after birth (through breastfeeding). A growing body of literature suggests that the intrauterine environment plays an important role for shaping future disease susceptibility in ways we do not fully understand. Fetal growth may not only affect organ size but also organ function in adulthood (Gluckman and Hanson 2005). The role of the uterine environment is therefore of great importance, but it is often difficult to disentangle its effects from those that are genetically mediated by the mother. The interaction between fetal and maternal genes might sometime trigger adverse outcomes, as may be the case for recurrent miscarriages. It should also be kept in mind that pregnancy can be life threatening for the mother: it is estimated that, in Africa, the lifetime risk for a woman to die of pregnancy-related causes is 1 in 16, while, in more developed countries, this figure is 1 in 1,800 or lower (AbouZahr 1998). Situations may thus occur when a pregnancy is interrupted for the sake of the mother's chance of reproducing again. Maternal, paternal, and fetal genes thus have different stakes in any given pregnancy (Haig 1993). Reproductive epidemiology has acquired a powerful tool with the new genetic technologies, and epigenetic changes (changes in gene function that occur without changes in the DNA sequence) in this time period may change gene expression. Epigenetics is a rapidly emerging area of research that may aid in the understanding of the impact of prenatal exposures on later health. These technologies will, hopefully, provide new important clues on the complex events taking place from conception to birth and onwards.

In epidemiological studies, we can generally only obtain data on pregnancies that end in delivery or miscarriage (and usually only on miscarriages of clinically recognized pregnancies, thus missing the early ones). As a consequence, most of the processes leading to events that result in early terminations are hidden or altogether absent from the data we have access to. We thus have to take into consideration what potentials for errors this entails. Access to ultrasound measures will change this in the future.

Especially for women, the time period of reproduction is relatively short and, in most countries, under intensive surveillance. Often, we can only study the causal links that remain after health-care providers have tried to prevent negative outcomes. Data on births are routinely collected in several countries at the national level and following standardized procedures, thus allowing comparisons over time and, in some instances, place. In some countries, data are computerized, facilitating access to researchers. Most countries regularly publish reports that include some indicators of reproductive health.

Reproductive health, at least the part associated with pregnancy and childbirth, covers events that can be frequent and serious, such as infertility, miscarriage, preterm birth, and pre-eclampsia. These are events we would like to prevent rather than treat, and most are willing to recognize that prevention must be based upon good quality research.

Not only do we need to identify determinants of reproductive failures connected to lifestyle, occupation, environment, diet, or medication but also to find out how antenatal care (ANC) can be best organized. Antenatal care is, in many countries, the most expensive part of preventive care within the health-care system, and much more research is still needed, especially in countries where resources are sparse. Much of this documentation has to be time- and place-specific. Evidence on health-care technology cannot just be transferred from one country to another. But many of the findings related to the etiology of given outcomes will apply to most populations.

By making efficient contraceptive methods more widely available, increasing the potential for ending unwanted pregnancies, and, especially, by developing more and more sophisticated methods for treating infertility, we are increasingly interfering with the forces of evolution and, to some extent, also with our ability to study reproductive epidemiology (e.g., how fertility declines with age, or to what extent the causes of infertility, rather than assisted reproductive technology (ART), affect pregnancy outcome). We should make it one of the priorities of reproductive epidemiology to study whether and to what extent ART will affect future generations. Reproduction is, in many ways, a playground for new technologies that are being introduced without much concern for the potential long-term consequences. Children born of ART (including in vitro fertilization (IVF), intracytoplasmic sperm injection (ICSI)) are still too young for us to study whether they may suffer any long-term health consequences associated with the mode of their conception, although these individuals are relatively young and we are thus still unable to evaluate properly whether their health has been affected. However, many of the studies that have been performed so far are reassuring. Interventions like ICSI may undermine the very core of the natural selective forces behind reproduction. At present we do not know if ICSI is a safe-assisted reproductive technology (Ludwig and Diedrich 2002), although early reports are reassuring to some extent about ICSI not being associated with more risks than IVF (Hansen et al. 2002; Lie et al. 2005; Steel and Sutcliffe 2009).

In the case of ART, it may turn out that trying to solve health problems for some may introduce health problems for their children. Some of these health problems may never come to the attention of physicians, and the potential consequences of infertility treatment must be brought to light by research. In relation to this and many other technologies, epidemiologists have an important role to play in monitoring disease data over time in the population. Reproductive health should be an important part of any health surveillance system.

It is peculiar to reproduction that, during pregnancy, the mother provides most of the exposures that may influence the future health of the unborn child. The fetus is only partly protected by the placental barrier. Many toxic substances pass this barrier, and the developing fetus will often be less able than the mother to metabolize these substances, especially early in pregnancy. This is, e.g., the case with alcohol, caffeine, and several drugs.

Thus, another peculiar feature of pregnancy is that not just one individual is involved in the process under study but three: the mother, the father (through fetal genes and, possibly, substances carried by semen (Ness and Grainger 2008; Savitz et al. 1994)), and the fetus itself through its continuing interaction with the mother during the pregnancy period. When studying reproductive outcomes, it is thus, in many cases, insufficient to focus only on the mother, although she usually represents the most accessible part.

It is now known that during pregnancy and delivery, there is a two-way traffic of cells between the mother and the fetus and that these cells can persist in the bloodstream for decades (Bianchi et al. 1996). This phenomenon, called "microchimerism," may interfere with the immune system and thus contribute to the etiology of several diseases, especially those of autoimmune origin (Bianchi 2000; Whitacre et al. 1999; Fugazzola et al. 2010). That pregnancy itself, or phenomena related to pregnancy, might have something to do with the etiology of these diseases is suggested by the fact that females are more subject to autoimmune diseases than males, although these differences may also depend on immunological, hormonal, or other factors that are related to sex but not specifically to pregnancy (Whitacre et al. 1999; Whitacre 2001). Changes during pregnancy may have health implications for the mother as well as for the child, and the growing fetus totally depends upon maternal supply of nutrition and oxygen. A growing area of research investigates whether changes in this supply may "reprogram" organ development, with possible long-term health consequences (Barker 1994; Gluckman and Hanson 2005). It is also well known that infections during pregnancy may affect the child. Torch was the acronym given to known fetotoxic agents (T = toxoplasma gondii, r = rubella, c = cytomegalovirus, h = herpes simplex, and o = the unknown agent). If infectious agents, or antibodies to infectious agents, attack the fetal brain, it could lead to neurological or psychiatric diseases that do not become manifest until much later in life.

45.1.3 Pregnancies as Repeated Events

Each pregnancy provides a new set of observations, and, in populations with a high fertility, a woman may be subject to repeated studies of exposures of interest for the outcome of each pregnancy. A pregnancy is a new event that, at the same time, shares many features with the previous pregnancy(ies). Most reproductive failures (such as miscarriages, preterm delivery, and pre-eclampsia) have a tendency to recur, probably because time-stable causes (e.g., genes or part of the maternal environment) are present during all events. However, it is not entirely clear how this aspect should be taken into consideration when analyzing data.

In animal studies, the variance within a litter and between litters is dealt with separately, probably even successfully, in the statistical analysis, but the situation is somewhat different for humans. Human births are usually sequential, except for multifetal pregnancies, which, however, are relatively rare and often not comparable to singleton pregnancies. Methods have been developed to take into account the statistical dependence resulting from women contributing more than one pregnancy (or one baby) (Hoffman et al. 2001; Watier et al. 1997), but these methods may not be able to properly take into account some of the issues specific to multiple pregnancies (Olsen and Andersen 1998). More recently, directed acyclic graphs (DAGs) have been used to illustrate how data from two pregnancies could be analyzed (Howards et al. 2007a; Olsen 2008).

Each new pregnancy has a number of features in common with the previous pregnancy and a number of features that are peculiar to it. In some situations, each new pregnancy is an independent event and should be treated as such, but the contrary argument can be made as well. However, even if the dependency between pregnancies were to be modeled in the analysis, it is not clear whether doing this would produce a different type of bias, because the statistical model will not take selective change in behaviors related to previous pregnancy outcomes into account. Pregnancies are, in many instances, planned as a function of the outcome of a previous pregnancy, and this aspect can hardly be addressed by methods aimed at accounting for the statistical dependency between pregnancies. A miscarriage or a stillbirth may be replaced very shortly with a new pregnancy in order to have the desired child. A surviving child with severe handicaps may delay or stop further childbearing, while parents of a child that dies with the same handicap may attempt a new pregnancy soon. Some women may avoid certain exposures as a result of how previous pregnancy ended. A mother who is a heavy smoker and had a child with a very low birth weight may well give up smoking when she becomes pregnant again, even in situations where smoking was not the only or even the major determinant of the low birth weight. If smoking was not the only cause of low birth weight, she may then become a non-smoking pregnant woman with a high a priori risk of getting a child with a low birth weight. How should this be addressed in the statistical analysis? Or how would we be able to design a proper counterfactual comparison? This type of confounding by the risk that triggers a change in exposure is very difficult to address, and it is a common and often neglected problem. A valid analytical solution may be to study first pregnancies only (Olsen 1994), but this approach is somewhat conservative and leads to loss of information, thus resulting in loss of statistical power. The occurrence of several outcomes changes with parity, in part due to selection (not all women will be able – or want – to have a new pregnancy) and in part, perhaps, due to physiology or changes in the uterine environment brought about by the previous pregnancy(ies). A first pregnancy may modify the uterus in ways that could facilitate the next pregnancy, as suggested by the finding that a long interval between pregnancies is associated with an elevated risk of pre-eclampsia in the next pregnancy (Basso et al. 2001; Skjaerven et al. 2002; Trogstad et al. 2001). Why this is the case is not known, but a pregnancy leads to both physical and endocrine adjustments that may modify risks in a subsequent pregnancy.

When studying the effect of a given determinant on an adverse reproductive outcome, one may be tempted to stratify on (or adjust for) a previous occurrence of that outcome (e.g., miscarriage). Such a temptation should be resisted if the intent is to study etiology, since adjustment for any factor that is caused in part by the exposure under study and is also correlated with the outcome will likely bias the estimate (Weinberg 1993), unless the objective of such a model is predictive and not epidemiological.

The fact that women tend to have more than one pregnancy does not, however, result only in difficulties and traps but provides reproductive epidemiologists with a powerful tool that is unique to this discipline. Many reproductive outcomes tend to recur in different pregnancies of the same woman, and this provides the opportunity to examine whether some putative factors play a role in the etiology of a given event by studying women who had the outcome in question in a pregnancy and estimating whether their recurrence risk changes as a function of a change in a given factor between the two pregnancies (Olsen et al. 1997; Basso 2007). This design, called "the computerized square dance design" (because square dancers continuously change partner) makes it possible to estimate whether the paternal genome plays a role by studying maternal half siblings. It can be applied to pregnancy outcomes as well as disorders occurring in early childhood, such as febrile seizures (Vestergaard et al. 2002). One of the advantages of this approach is that a great deal of confounding is adjusted for by using the woman as her own control. The time between pregnancies may itself play a role in the occurrence of the outcome of interest, and thus the interval between pregnancies may have to be taken into account, especially when studying change of partners, since women who change partner tend to have a much longer interpregnancy interval (Basso et al. 2001; Skjaerven et al. 2002).

45.1.3.1 The Problem of Incomplete Denominators

Ideally, we would like to be able to study the outcome of all conceptions that take place in a given population in a given time period, in order to be able to observe how many end in very early losses, how was the karyotype of the lost fetuses, their sex ratio, etc. Less ideally, we would like to be able to obtain information on all conceptions surviving the first 8 weeks of gestation, because missing them may lead to serious bias in studies examining specific exposures. In many instances, reproductive epidemiologists work with less well-defined cohorts and are thus often the victims of incomplete denominators.

45.1.3.2 Miscarriage

A frequent adverse pregnancy outcome, miscarriage, is to a large extent part of nature's own quality control system (Quenby et al. 2002). Some miscarriages are, however, the result of external exposures and could, in principle, be avoided if their causes were identified. A large number of losses occur in the preclinical phase, before the pregnancy is recognized (Wilcox et al. 1988), while others occur after the pregnancy is recognized. The timing of observation thus becomes of crucial importance and a potential source of bias.

The best approach to study fetal losses would be to start observation before conception (even though a pregnancy can only be recognized after implantation), but this requires access to women who plan their pregnancies and are willing to provide daily urine samples (Wilcox et al. 1988). Pregnancy planners are likely to include more subfecund women (women with a low probability per cycle of conceiving), as women who become pregnant as a result of contraceptive failures have, generally, higher fecundity. Participating women may thus also be at an elevated risk of a number of adverse outcomes, since subfecundity appears to be correlated with several adverse pregnancy outcomes (Henriksen et al. 1997; Draper et al. 1999; Basso and Baird 2003; Basso et al. 2003; Zhu et al. 2006, 2007a). To reduce the proportion of subfecund couples taking part in a study, one could enroll only couples who discontinue contraception in order to have a child or who have been trying for no longer than six cycles. This requires screening of a large number of couples but will lead to a study sample that is likely more representative of the underlying fertility of the population of interest (Baird et al. 1986; Weinberg et al. 1993, 1994a).

Currently, few investigators have attempted to detect pregnancies by using biomarkers (usually hCG, human chorionic gonadotropin). These studies (Wilcox et al. 1988; Bonde et al. 1998a, b) do not tell us how frequent miscarriages really are, but they suggest that at least 30% of conceptions end as miscarriage and that a little more than half of these occur in the preclinical phase.

An exposure that delays miscarriages without increasing their incidence would appear as a risk factor in a study based exclusively upon recognized pregnancies. Such an exposure may move a loss from the preclinical phase to the detectable phase. In contrast, an exposure that advances the time of loss would appear to prevent miscarriages. Such an exposure would produce a low miscarriage rate because miscarriages would tend to occur in the preclinical phase among the exposed. For this reason, the timing of pregnancy diagnosing is important. Women who take a long time to conceive may be more aware of early losses than women who have not waited a long time if the former tend to seek earlier confirmation of pregnancy. This type of bias may be partly responsible for the association between subfecundity and miscarriage (Baird et al. 1993). The use of early pregnancy tests may play a role in analyzing and interpreting data from studies based upon pregnancy planners. Such studies are, however, difficult and expensive to conduct and require a very cooperative study population. If women are recruited at different times during pregnancy, the statistical analysis should account for time of entry to avoid bias (Howards et al. 2007b), especially in circumstances where the exposure may in part determine when women are recruited.

When studying environmental determinants of miscarriage, you may want to exclude the so-called habitual aborters (a term generally applied to women who have miscarried three or more pregnancies). Such women will not provide information related to risk following the exposure (Gladen 1986; Weinberg et al. 1994b). The problem is that these particular women cannot be identified by their history of loss alone and cannot be identified at all if they have no pregnancy history. Stratifying results on pregnancy number can distort associations, since only some of the women with several miscarriages will be habitual aborters. Some miscarriages will be due to chance, while others could be caused by the exposure under study.

45.1.3.3 Measuring Infertility

When we are interested in measuring the biological component of fertility, called fecundity, we are often able to ask pregnant women how long it took them to conceive. Couples who had planned to become pregnant will provide the best-quality information, but difficulties are attached to the fact that “planning” is often not a well-defined concept. Furthermore, some respondents may not have planned a pregnancy but will nonetheless report that the pregnancy was planned when asked. As mentioned before, among planners, there will be a number of subfecund couples, and therefore we should also find a way to study couples who had not planned their pregnancy and those who became pregnant despite using contraception, as their underlying fecundity may differ from that of couples who plan a pregnancy. We would also like to have information on all couples that conceived and had an early pregnancy loss (Jensen et al. 1998) or failed to conceive because they were sterile, and we would also like information on those who gave up trying (Bonde et al. 2006). When studying time to pregnancy in samples of pregnant women, one must be reasonably confident that giving up a pregnancy attempt is completely independent of the putative risk factors under study (Basso et al. 2001), a condition similar to what we encounter when working with censored data in general.

45.1.3.4 Congenital Malformations

Congenital malformations are relatively frequent (2–6%), and they constitute a major cause of infant mortality and morbidity. Having a child with a congenital malformation is also a very stressful event for the involved family, which has led to concerns about the possibility of obtaining comparable information in a case-control study addressing determinants of congenital malformation.

It is now generally accepted that the proportion of congenital malformations at birth is a measure of prevalence. The incidence of congenital malformations is not available for study, given that it requires counting of new events since time of conception. Congenital malformations that occur in utero and end as miscarriages, some as very early miscarriages, are usually not detected. Some exposures may simultaneously increase the incidence of some congenital malformations while at the same time reducing the survival rate of the affected fetuses (or embryos) in utero, thus decreasing the prevalence at birth. Monitoring systems of congenital malformations with no data on miscarriages or terminations due to the diagnosis of a malformation could therefore miss important teratogens and even wrongly conclude that a given factor is protective. Many of the established monitoring systems on possible teratogenic effects of medicines taken during pregnancy are based only on data on prevalence at birth. When dealing with miscarriages and congenital malformations, missing early losses may lead us to biased effects measures. Many of the routine monitoring systems are furthermore of poor quality, partly because only some of the congenital malformations are visible at birth. Some heart defects may, e.g., only produce symptoms – and thus first surface to clinical detection – under extreme physical strain late in life (Knox et al. 1984).

45.1.3.5 Time Matters

If you were a student of mortality, you would know that the question is not *if* people die but *when* they die. Mortality rates (*MR*) reflect this time function, as $1/MR$ is the life expectancy in the same time unit as the rate is measured, given that a number of strong conditions are fulfilled. All estimates of risks come with a time tag. The estimate depends upon the length of time it represents. Time is underlying all occurrence research, even when it is not explicitly mentioned since all events happen in time. If you do not wake up in the morning, it may be because you are dead. You ran out of time.

In reproductive health, time is important from several points of view, not only as the time from exposure to the endpoint of interest. The timing of exposure itself is more important in this area than in most others. Specific windows of vulnerability open and close over the time of gestation; the time of organogenesis and organ development plays a crucial role for congenital malformations and many other outcomes. Not only may malformations of genital organs impair reproduction, but the number of Sertoli cells is at least partly determined in fetal life. Exposures that reduce Sertoli cell production may have implications for sperm production in adult life (Sharpe and Skakkebaek 1993; Wilcox et al. 1995). Organ development is under the influence of hormonal factors that operate at certain time periods. Fetotoxic exposures may have different outcomes as a function of the timing of exposure, and growth-determining factors may only play a role during a time period of rapid fetal growth.

Early losses appear to account for the largest proportion of losses (Wilcox et al. 1988). Although we have very little direct knowledge about the determinants of these early losses, many are likely due to chromosomal abnormalities or other defects. Among clinical losses, there is an excess of congenital abnormalities compared with birth (Shepard et al. 1989). Identifying avoidable losses depends to a great extent on the timing of the exposure which may be difficult to identify, partly because fetal death need not coincide with the expulsion of fetal tissue. Fever could, e.g., appear to be a cause of miscarriage, because fever may be induced by dead fetal tissue. In that case, the association would be from fetal death to fever and not from fever to fetal death (reverse causation). However, Andersen et al. (2002) did not find that routine episodes of fever were associated with miscarriage.

The period at risk for teratogenic actions (usually the second and third months of gestation) is almost over when antenatal care (ANC) begins. All actions taken to reduce, e.g., occupational exposures during pregnancy are usually put into operation too late. Some types of prevention (like the use of folic acid to prevent neural tube defects) need to be started even before the pregnancy is planned. If a given toxicant (such as lead) is stored in the maternal tissue over a long period of time, its mobilization during pregnancy may affect the mother or the embryo many years after the exposure took place (Rothenberg et al. 2002).

It is likely that relevant exposure windows for specific effects are only open during short time periods. Exposure to influenza virus has, e.g., been associated with schizophrenia in the offspring, but only for infections during the second trimester of pregnancy (Mednick et al. 1994). On the other hand, neurotoxic exposures may influence brain development, if they operate at the time of the vulnerable structural changes that take place throughout fetal life and in early childhood (Sun et al. 2009). It is possible that diseases like ADHD (attention deficit and hyperactivity disorder), autism, or Tourette's syndrome have a pre- or perinatal etiology.

The nutritional demand of the fetus is at its maximum in late gestation. Smoking in this period is much more closely associated with low birth weight than smoking early in pregnancy (Royal College of Physicians of London 1977).

The Dutch famine study suggested that the timing of undernutrition during gestation may play an important role for later health outcomes (Susser and Stein 1994). Being exposed to undernutrition in the last trimester was associated with obesity later in life (Ravelli et al. 1976).

Most side effects of drugs taken during pregnancy depend on the timing of drug intake within the gestational period. This is likely true not only for teratogenic effects but also for more subtle outcomes, such as learning disabilities or developmental milestones (Zhu et al. 2009). Drugs are particularly difficult to study as an exposure, because it is difficult to disentangle the effect of the drugs from the effect of the indication for taking the drug in the first place (what is usually termed "confounding by indication") (Olsen et al. 2002). If two or more drugs with the same indication but different active molecules are available and are prescribed to pregnant women independently of the severity of the indication, this may provide useful options for better characterizing the effects related to one specific drug.

The fetal alcohol syndrome (FAS) is characterized by specific facial characteristics, low birth weight, and cognitive impairment. It is likely that these impairments are the result of time-specific exposures and, if so, binge drinking may be of concern (Sun et al. 2009), even binge drinking among low or moderate alcohol drinkers. Still, no studies in humans have shown that binge drinking alone causes cognitive impairments or any other important FAS characteristics, perhaps because all studies have been too small. However, binge drinking has been associated with an increased risk of cleft defects in a recent study (DeRoo et al. 2008). If the exposure window for binge drinking is short, then only a small fraction of binge drinkers will be at risk at any given time, and small studies will thus have limited power to detect an effect. FAS is seen in children of some women, who drink large amounts of alcohol every day. A high daily intake leads to exposure during the time periods of vulnerability, regardless of their duration.

Exposure to mercury, determined on the basis of amounts found in umbilical cord blood at birth, has been associated with impaired cognitive functions (Grandjean et al. 1997). Measurement of mercury in cord blood reflects an average exposure throughout gestation, and, as such, it does not indicate whether certain time periods are more vulnerable than others.

Hormonal factors play a profound role in fetal life. Sexual hormones are, e.g., needed for a fetus to develop a male phenotype. It is furthermore believed that hormonal factors influence the probability of developing some diseases in adult life. For example, it has been suggested that a high intrauterine estrogen level may modify breast cancer risk (Trichopoulos 1990) and reduce the number of Sertoli cells (Sharpe and Skakkebaek 1993), thus resulting in an increased risk of undescended testis, hypospadia and testis cancer, as well as reduced semen production in adulthood. This latter hypothesis is, however, not corroborated by epidemiological findings in general, and it probably has to be modified to fit existing data (Strohsnitter et al. 2001; Storgaard et al. 2006). There is, on the other hand, clear evidence of an association between some of the above diseases. Patients with testis cancer have a higher frequency of undescended testis and apparently a lower fecundity before the cancer is diagnosed (Jacobsen et al. 2000). Macro-epidemiological (ecological) studies furthermore show that the geographical variation in, e.g., testis cancer is often followed by similar geographical variations in low sperm count and a high prevalence of malformations of the male genitalia. The link between testicular cancer and undescended testis is relatively well established, although most studies have been based upon self-report of undescended testis, often reported retrospectively. In many instances, undescended testis descend spontaneously shortly after birth or at puberty. More recent studies based upon recording at birth show a weaker association between testis cancer and undescended testis than that reported in studies based on recall (Stang et al. 2001; Sabroe and Olsen 1998; Trabert et al. 2013). It is possible that only persistent undescended testis correlate with the risk of testicular cancer. It is unclear whether descent through treatment eliminates the risk or whether the risk is caused by the underlying reasons of the displacement itself. If the risk is a function of underlying hormonal disturbances at the time period of organogenesis, treatment is not expected to affect cancer risk.

Single peak exposures to drugs are seen in pregnant women who try to commit suicide by taking an overdose of medicine. In principle, these studies provide unique data for studying the fetotoxic effects of these drugs, since they often bypass confounding by indication (unless the study concerns anti-depressant drugs). Many of these studies, however, have not taken abortions (induced or spontaneous) into consideration and have often been too small to detect teratogenic effects, since such studies must focus upon suicide attempts in specific and short time periods. Furthermore, many women terminate their pregnancy after a suicide attempt because of fear of a fetotoxic effect or because of their inability to cope with a pregnancy. Newborn children available for this type of study are thus few and selected. Most of the findings from these studies have been reassuring, indicating that healthy babies are often born after a single exposure to a high dose of a specific drug (Flint et al. 2002).

Although it is generally accepted that timing of exposure is of crucial importance for most outcomes of pregnancy, many studies rely on information where the timing of exposure is not specified. Many studies make use of retrospective data from pregnancies and exposures that are difficult to recall, as they have taken place years before reporting. Large cohort studies have been started or are being planned in order to provide researchers with prospective data on exposures (Olsen et al. 2001; Birthcohorts.net 2011).

45.2 The Case-Parent Triad Design

There are several design options for studying genetic risk factors. Many of these are described in chapters ▸Modern Epidemiological Study Designs and ▸Statistical Methods in Genetic Epidemiology of this handbook. One of these designs is, however, of particular interest in reproductive epidemiology, namely, the case-parent triad design. The idea is to use parents of affected children to serve as genetic controls. The maternal genome can be very relevant to early outcomes, such as birth defects, and case-control designs that do not study the mother may result in the discovery of apparent genetic effects in the offspring that are really just confounded with maternal genetic effects. Case-control studies may also be confounded by genetic factors when they are applied to populations with a mixture of different ethnic groups. The parents are by definition ethnically matched to the case, and they are furthermore usually motivated to take part in the study. The two non-transmitted parental alleles can be compared to the two transmitted alleles, and since this Mendelian transmission occurs at random, the observed allele structure can be compared with the expected values. If the parents, e.g., both have the allele structure Aa, the children will under Mendelian transmission be AA, Aa, Aa, or aa with equal probabilities, 25% probability of AA and aa, respectively, and 50% of Aa. If cases with such parents differ from this, it may be because these genes are causally linked in the disease.

If the genotype is associated with the disease under study, affected children will have an allele distribution that deviates from expected values according to Mendelian principles. All this may be analyzed using well-known log-linear models (Weinberg et al. 1998). The main limitation in using this design for diseases with an adult onset is that the assumption of non-selective survival of the parents to the time of the study is strong, at least for severe diseases. For diseases with an early onset, the assumption will usually be fulfilled. Using this design to, e.g., study genetic causes of congenital malformations is an attractive option. The parents will be available, motivated, and present in a setting where a saliva or blood sample can easily be taken at any time.

45.3 Infertility and Subfecundity

45.3.1 Measures Used to Describe Fertility

Fertility refers to the reproductive capacity of an individual, although there may be differences in meaning, as explained later in this section, depending on whether the term is used in demography or in clinical settings. The fertility rate is defined as the number of live births a woman has during her reproductive life. A fertility rate of 1.3 means that in a population of 100 women, there will be, on average, 130 liveborn children – not enough to replace their parents if migration in and out of the country is in balance at all age groups. A fertility rate of more than 2.0 usually is needed to keep the population in a steady state (as slightly fewer females than males are born and some will die before they reach the age of reproduction). How much more depends upon the life expectancy and the sex ratio of the newborn children in the population, which generally varies slightly but can be severely altered in countries where a strong preference for a specific sex exists. On average, a woman should produce one girl who survives until the end of her reproduction age in order to keep the population in a steady state, if life expectancy is stable over time.

Fertility is determined by both a biological capacity to reproduce (fecundity) and the desire for a given family size, which is under the influence of social and cultural conditions, broadly defined. Available methods for family planning play, of course, a crucial role and, in many ways, interfere with the epidemiologists' ability to study fertility, as many couples plan their reproduction as a function of many factors, including their professional ambitions and their past reproductive history.

Fertility is a term used by demographers to describe the actual production of live children, whereas infertility is used by the medical profession to describe a reduced biological capacity to reproduce. This dual use of the same terminology is confusing.

We suggest that the term fecundity be used to describe the biological capacity to reproduce and fecundability (the probability of conceiving within a given menstrual cycle) to be a quantitative estimate of fecundity. Unlike Cramer and Goldman (1994), we propose to let the more specific details of the terms be defined in the actual study, in order not to have too many words describing rather similar conditions. Reproduction, at least historically, requires sexual contact, fertilization, implantation, and survival to birth and beyond. Now alternative methods exist.

Most studies use a recognized pregnancy (or birth) as the endpoint, and recognition of pregnancy usually depends upon clinical or biochemical measures. Fecundity could, therefore, depending on the specific study, describe the ability to obtain a biochemically detectable pregnancy (by means of hCG), a clinically recognized pregnancy, or a pregnancy that led to a liveborn child. Fecundity can thus be used to describe the capacity to achieve any of these endpoints, which may be confusing. On the other hand, having a specific term for each of these situations would complicate communication to an even larger extent. We will use the term fecundity to refer to a variety of situations.

Childlessness can be voluntary or involuntary, and the latter may be subject to epidemiological research. Subfecundity (i.e., taking a long time to conceive while having unprotected sexual intercourse) is a frequent problem, and treatment for subfecundity is a rapidly growing sector of many health-care systems. Research that can potentially lead to prevention of subfecundity is therefore receiving increasing attention, and epidemiology plays an important role in this research. Subfecundity is a measure for a couple, defined on the basis of unsuccessful attempts for a given length of time (often set at 6, 12, or 24 months). In many industrialized countries, about 15% of all couples that at least try to become pregnant will experience subfecundity, if this is defined as a waiting time of 12 months or more (Juul et al. 1999). The term infertility usually describes an unsuccessful waiting time of 12 or 24 months. Often the cut-point of 12 months is used in affluent societies and 24 months in countries with limited health-care resources because the term provides access to treatment in some countries.

Sterility is defined as an absent capacity to reproduce – a fecundability of 0. Since most couples' probability of conceiving is very rarely 0, many women will eventually become pregnant if they keep trying. Some of the women who become pregnant after infertility treatment would have become pregnant without receiving the treatment.

Time to Pregnancy (TTP) If a normal fecundability is 0.25, then 3% of couples with normal fecundity will not succeed within 12 months of trying ($(1\text{–}0.25)^{12}$) just because of bad luck. These couples need no treatment but may (and some will) be treated nonetheless, usually with an "excellent prognosis" (25% success rate for each cycle). The remaining 12% who had a TTP of 12 months or more are the proper candidates for treatment. If all couples who wait unsuccessfully for 12 months or more to become pregnant receive treatment, then any treatment will to some degree be successful. Some couples will have normal fecundity, and some will have subnormal fecundity without being sterile. An effective treatment has to demonstrate better performance than chance alone would produce. This is not always easy to measure, as many treatments involve transfer of more than 1 embryo, which will result in higher chances of a live birth than if only one embryo was implanted.

If a couple is defined as subfecund after an unsuccessful waiting time of 24 months, then only about 0.1% of "normal" couples will meet the definition, and those targeted for treatment will include 99.9% of actually subfecund (and sterile) couples. For this reason, less affluent societies do not start infertility treatment until couples have tried for at least 24 months.

Since fecundability is related to the duration of a waiting time to pregnancy (TTP), TTP is a frequently used measure in subfecundity research. The measure was first used by demographers when they examined the time from marriage to the first liveborn child; in this context, TTP is a measure that only has biological relevance in societies where procreation starts with marriage and contraception is not practiced in that time period.

In societies with a higher degree of family planning, TTP (now defined as the number of cycles – often approximated by months – that elapse from the time a couple first starts trying to when they actually conceive a clinically recognized pregnancy) becomes a useful tool in the study of determinants of subfecundity. It has been used in epidemiology since the early 1980s (Rachootin and Olsen 1982, 1983; Olsen et al. 1983; Baird et al. 1986). It appears that women are able to recall with sufficient accuracy how long they took to conceive, even after several years (Joffe et al. 1993, 1995), although there are also indications that accuracy of recall may not be very good (Cooney et al. 2009). Time to pregnancy is easy to use and, perhaps for this very reason, has undoubtedly been used too frequently without proper concern for its shortcomings and pitfalls, which have been described in detail (Baird et al. 1986; Weinberg et al. 1993, 1994a; Basso et al. 2000a; Olsen and Rachootin 2003). When comparing TTP values in different populations, you should take into consideration frequency of sexual intercourse, the desired family size, and the persistence in trying to become pregnant at least for some time. And, as always, you have to take into account the potential confounders of the association of interest.

45.3.2 Design Options in Studies of TTP

The best option to study fecundity would be to follow women from when they first become sexually active, recording contraceptive use, intercourse, pregnancy, etc. However, this is hardly feasible. The next best option is to study TTP in a longitudinal design that starts when couples stop using contraception in order to conceive (starting time). Exposures of interest can then be recorded at the relevant point in time (and independently of the length of TTP). Generally, the starting time (when couples start trying to conceive) is the time when exposures should be recorded, although if prior information on exposures exists, that can be used. Exposures should also be registered for attempts that do not lead to a pregnancy. Studies that rely on exposures recorded at the time of pregnancy rather than the starting time may produce biased results, especially for exposures suspected to cause infertility. For example, if smokers stop smoking after having tried in vain to become pregnant for a certain time period, but they do not quit if they conceive quickly, smoking recorded at the time of conception will correlate with a short waiting time, as if smoking prevented subfecundity. In fact, abundant evidence points towards the opposite assessment: smoking impairs fecundity, at least in women (Rachootin and Olsen 1983; Baird and Wilcox 1985; Bolumar et al. 1996; Alderete et al. 1995). However, other sources of bias may contribute to exaggerating the effect of smoking or other exposures on fecundability. Studies restricted to couples who plan their pregnancy, as they can more reliably report a time to pregnancy, do not include couples with unplanned pregnancies. If an exposure is associated with a higher risk of having an unplanned pregnancy, the most fertile among the exposed will enter the study when they start planning, while those who have conceived by accident (usually the more fertile) will be excluded. This way, the effect of an exposure may be exaggerated in studies restricted to planners (Weinberg et al. 1994a).

The proportion of couples that become pregnant during the first cycle is an estimate of the fecundability rate in the population. The entire TTP distribution will, however, normally be used in the analysis, e.g., in a discrete Cox model. Such a study need not last 12 or 24 months. Even a study with a follow-up time of only one cycle could provide evidence that some exposures have an effect on fecundability. If the estimated effect is a 20% decrease in fecundability, we would expect 30 out of 100 non-exposed women to be pregnant within the first cycle versus 24 out of 100 exposed women. Such a study of pregnancy planners is straightforward to design and to analyze but extremely difficult and expensive to carry out, unless it is based upon couples that are seeking treatment for infertility, like IVF patients (Olsen et al. 2005). Results from infertility patients are, however, hampered by limited generalizability to the population at large, in part because the couples are highly selected, in part because spermatozoa, eggs, and embryos are selected by health personnel according to criteria that may be quite different from those of "natural" selection. A longitudinal study on pregnancy planners may also provide other related endpoints, such as information on early losses or semen quality.

Epidemiologists have looked for less expensive designs than the concurrent follow-up of pregnancy planners. These designs have mainly been population-based cross-sectional studies or designs based upon samples of pregnant women, both of which are prone to a number of problems, although they are less expensive and easier to carry out.

Cross-sectional surveys rest upon the assumption that women recall their TTP with some accuracy, even a long time after the event. In a survey, the selected women will try to record attempts to become pregnant, exposures of relevance – often from the same time period – infertility treatment, and TTP (or time waited in unsuccessful pregnancy attempts). Population studies of this type are relatively inexpensive, and they usually rest on random sampling principles. The main problem is inaccurate recall, especially of exposures at the relevant point in time (the starting time), and the ability to obtain response rates that do not introduce selection bias. Analyzing data may be simple, although TTPs usually have to be recorded in rather broad categories, often referring to months rather than to cycles. Women may be able to recall if they had to wait for at least 6, 12, or more months to become pregnant, but usually they will not be able to remember if they became pregnant after 5, 6, or 7 months of waiting time, when the relevant event took place several years before (although women will probably remember if they became pregnant in the first month of trying). Digit preference also has to be taken into consideration, as clusters of reporting will be seen at specific waiting times (such as 6 or 12 months).

Using pregnant women in data collections has, for obvious reasons, been a convenient design. In most countries, pregnant women are easy to locate and to contact. They are usually more willing to take part in studies than other women, and they are generally able to accurately report their TTP if the pregnancy was planned. Women usually remember when they started planning a pregnancy, and may be able to recall the exposures around the starting time, at least if the planning did not start too long ago. The way questions about planning a pregnancy are asked is important, because the notion of having planned a pregnancy is open to interpretation, and its desirability may increase once a woman is pregnant. The question should be as specific as possible and at the same time formulated in such a manner as to elicit the most honest response.

Since pregnancy is a condition for participation in this type of study, this design will result in the exclusion of sterile couples and is thus unable to reveal an all-or-none exposure effect (i.e., an exposure that either causes sterility or does not delay conception). However, most exposures that we know of are not of this type, and the exposures that reduce fecundability will show a longer TTP in a pregnancy-based study. The qualitative effect measure that comes from a study of this kind is, however, not a measure of fecundability in itself (Olsen and Andersen 1999), although it will correlate with fecundability under a number of conditions. The most important, and often neglected, of these conditions is the persistency in trying to become pregnant. This persistency should not be related to the exposure under study, because, if so, the exposure may falsely appear to increase fecundity (Basso et al. 2000b). The timing when diagnosing a pregnancy also plays a role. If couples wait very long to have a pregnancy verified, then an early loss may count as waiting time. Had women used a sensitive pregnancy test early, the event would have been recognized as a TTP of a given length followed by an early loss. The length of the waiting time may influence the timing of detection, thus potentially producing bias (Baird et al. 1993).

Exposures that cause irregular or long cycles may also interfere with the timing of pregnancy recognition as well as with the woman's ability to report the starting time and the waiting time to pregnancy.

The use of family planning methods should be taken into consideration. Couples who use unsafe contraceptive methods without getting pregnant will, over time, include an increasing fraction of couples with a low fecundability. Couples who know how to time sexual intercourse around the time of ovulation have shorter waiting times than other couples, all other factors being equal. Couples who, in the past, have experienced subfecundity may have since modified their exposures or sexual behavior in a way that may impair our ability to find a proper reference group.

Starr and Levine (1983) have suggested the calculation of a standardized fertility ratio. The underlying idea is to compare the observed fertility measured by liveborn children with the expected fertility for couples of the same age and from the same region. Such a standardized fertility ratio can then be calculated before and after an exposure of interest. If the exposure impairs fecundity, the standardized fertility ratio after the exposure divided by the standardized fertility ratio before the exposure would be less than one, other things being equal. The observed number of liveborn babies before exposure may be close to expected values based upon age and calendar time-specific rates in the population at large, but the exposure may then reduce this observed/expected ratio. Although this method was able to detect the fecundity-reducing effect of dibromochloropropane (DBCP) (Levine et al. 1981), it rests upon strong assumptions, and it only works when examining exposures that have a specific (and known) starting point in time.

In addition to the sources of bias mentioned here, there are numerous other sources of bias in time to pregnancy studies (Weinberg et al. 1994a; Weinberg and Wilcox 2008; Baird et al. 1994). It is not an easy task to develop a monitoring design sensitive enough to pick up subtle changes in fecundity over time (Olsen and Rachootin 2003), which is unfortunate. We would like to detect changes in fecundity that may correlate with widespread environmental pollutants, such as perfluorinated compounds (PFCs), which may be associated with decreased fecundity (Fei et al. 2009).

Only about half or less of those who experience infertility seek medical help (Olsen et al. 1998a, b), making patients of infertility clinics a selected sample of infertile couples. The factors that determine which couples seek treatment should therefore be considered when using infertility patients in epidemiological studies.

If the forces of selection are correlated with the exposure under study, then a case-control study with cases defined by treatment is not a valid option for an analytical study. We expect these forces of selection to treatment to be related to several factors, such as cultural background of the population, age, parity, education, social status, and availability of treatment facilities. Some of these conditions may well correlate with lifestyle factors, dietary factors, infections, or other putative causes of subfecundity.

Given these conditions, the source population has to be defined by the case series. The source population could be defined by all those who would come to the treatment center if they had a similar infertility problem as our cases had in the time period of study. The source population is then defined by all potential cases (and, of course, by the enrolled cases as well). Had we known the source population, we could have sampled controls from this group at the time the cases came to be detected. Candidates for controls selection would then be couples who have planned a pregnancy and would be cases if they had a waiting time of 12 months or longer. We could then compare exposures at starting time, or before, and estimate the relative risk of being infertile as a function of the exposures under study. The problem in this design is that we do not know the source population and it may be impossible, even in principle, to identify it along with its exposure experience. A viable option may be to take advantage of the fact that infertility is a couple phenomenon. Infertility is sometimes caused by exposures that affect females, sometimes males, and sometimes both males and females. If we take an interest in exposures to prenatal tobacco smoking as a cause of "male infertility," for example, we may then compare the frequency of this exposure for males in couples that had a male problem (e.g., poor sperm quality) with the exposure frequency in males from couples where the female was identified as having the problem. Both sets of couples sought help and therefore belong to the source population. Since couples where both members have a fecundity problem belong to both the case and the control group, they do not provide useful information for the question under study and can be excluded. The potential problem in this design is that, often, both individuals in a couple seeking help for infertility have low fecundability, even if a specific problem is only identified in one member of the couple. Also, it is not guaranteed that all possible causes of infertility are investigated in both partners, as several may be present.

Alternative design options for monitoring fecundity have been proposed like using dizygotic twinning rates as a surrogate for fecundity (Tong et al. 1997), but one of the problems is to exclude twins that are a result of infertility treatment.

Studies on semen quality may also be used as a surrogate measure of fecundity. Much of the concern for a declining fecundity stems from studies on semen (Carlsen et al. 1992), although the quality of these comparisons over time of semen studies is questionable at best.

The main problems related to using measures of semen quality in epidemiological studies stem from low participation rates and difficulties with obtaining comparable conditions in the analyses. Several factors have to be taken into account: time since last ejaculation, the conditions related to the ejaculation itself, the time from ejaculation to analysis, the technical conditions for the analysis, the season of sampling (as sperm counts are higher in winter than summer periods in temperate climates), and all other potential confounders. Specific diseases may interfere with sperm production, together with external exposures.

45.4 Twins

Dizygotic twins have been seen for a long time as an indication of high fecundity, given that they require two eggs to be fertilized by two spermatozoa within a short time period. However, only recently have studies shown a direct link between time to pregnancy and twinning (Basso et al. 2004a; Ferrari et al. 2007; Zhu et al. 2007b). The rate of dizygotic twins varies geographically and by ethnicity, and mother's height and body mass index may have an influence (Basso et al. 2004b; Reddy et al. 2005). The "natural" incidence in Caucasian populations is about 1 in 80 births. Monozygotic twin rates are, on the other hand, more stable. For a time, a decline in dizygotic twins in several countries was, by some, interpreted as a decline in fecundity across several populations (James 1982, 1986), although the mechanisms responsible for such putative decline have not been firmly identified, and twinning rates started increasing again. Whether this increase is only the result of an increasing use of Clomid and IVF or whether the natural twinning rate also increased is difficult to determine, as birth records rarely reflect whether a birth was secondary to IVF or Clomid.

Twinning is associated with a higher risk of perinatal morbidity, including congenital malformations, and an increasing number of countries with health care providing IVF cycles to infertile couples are recommending single embryo transfer, to avoid increasing the rate of twinning.

Twins provide a very interesting source of data concerning the nature-nurture discussion in disease occurrence. Twin studies have been the basis for studies that tried to disentangle the effects of genes and those of the environment.

Genetic disorders are expected to have higher correlation (concordance) in monozygotic (MZ) twins than in dizygotic (DZ) twins. MZ twins are genetically identical (although some minor genetic differences have been found to exist and have even been used to identify a genetic cause of a congenital malfunction (Kondo et al. 2002)), and DZ twins share genes like ordinary sisters/brothers. Unlike ordinary siblings, however, DZ twins share the same uterine environment as well as very similar conditions in early childhood. These factors considerably reduce the confounding that would arise by comparing ordinary siblings. Although the twin model is definitely of interest, the design does not allow the quantification of heritability, as was previously believed. MZ twins have the same gene map but need not have the same functional genetic expression. Since females are mosaics where a random X chromosome is inactivated in each cell line, female twins are not functionally genetically identical. Epigenetic characteristics and genetic imprinting may differ between MZ twins (boys and girls). Furthermore, MZ twins' intrauterine conditions may differ, in terms of chorionicity or transfusion syndrome, for example, which results in often large variation in size between babies in a pair. Finally, both twins have to survive to be part of a twin study, and twin mortality is higher than in singletons, most likely from the time of conception, thus resulting in selected pairs of twins being born. It is well documented that a number of pregnancies start as twins, but both fetuses do not always survive, and it has been suggested that the surviving twin is at increased risk of cerebral palsy (Pharoah and Adi 2000).

Because twins are relatively rare and differ from singletons in terms of birth weight, gestational age at birth, mortality, and morbidity risks, they are often not included in epidemiological studies on potential fetotoxic hazards.

Twin pregnancies have been used as a model to study the effect of hormone exposure during pregnancy. The estrogen level is high in twin pregnancies, and twins of different sex also offer unique intrauterine exposure conditions (Storgaard et al. 2002). One can use opposite sex DZ twins to study differences in sex hormone exposures during prenatal life. It is, for example, believed that DZ females that have a twin brother have been exposed prenatally to a higher testosterone level than DZ females that had a twin sister, and some believe that this may affect their fertility (Lummaa et al. 2007), although some evidence suggests no differences (Christensen et al. 1998). It has also been suggested that a high intrauterine exposure to testosterone may increase the risk of autism in childhood (McClure 2003).

45.5 Measuring Adverse Reproductive Outcomes

Estimating the frequency of adverse reproductive outcomes is, in several ways, similar to estimating the frequency of any other disease. Frequencies are measured by means of rates or proportions (and many proportions are, unfortunately, called rates). Incidence measures new events over time, while prevalence describes an existing state at a given point in time.

Rates are used to present reproductive failures as a function of time, e.g., the incidence of cervical cancer in Danish women in 1997 was 11.42 per 100,000 person-years. The cumulative risk is estimated by the proportion of people who contract a specified illness over a given time span. Estimation can be made directly in a population where follow-up is complete: for example, the cumulative risk of stillbirth can be estimated if a sufficiently large number of pregnant women are followed from their 24th week of gestation until they give birth (if 24 weeks is the threshold used to distinguish between miscarriages and fetal deaths). Rates describe occurrence of events per time unit, like new respiratory syncytial virus (RSV) cases per 1,000 observation months in children less than 1 year in a given region and a given time period. Rates thus estimate disease occurrence in populations. Risks indicate estimates of the probability of a given event in a given time period for an average member of the population in question. In a population with complete follow-up, the risk of, e.g., second-trimester miscarriage in 500 pregnant women who are followed until the start of the third trimester would be estimated as 0.02 if 10 of the 500 women experience a miscarriage in these 3 months.

Incidence rates and cumulative risks are based upon data from a population at risk, that is, the population at risk of getting the disease (which excludes individuals who already have the disease when follow-up starts). Time is generally included in the measure. Prevalence proportions are estimated as the number of people with the disease in question at a given point in time divided by all in the population at that time, regardless of their disease status. If 500 women are pregnant in a population of 50,000, the pregnancy prevalence proportion is 0.01 (500/50,000). Since prevalence (P) is a function of incidence (I) and duration (D), the incidence of new pregnancies in a population in steady state would be 500/8, if the average duration of a pregnancy is set at 8 months (to take miscarriages into consideration), that is, 62.5 pregnant women per month in the population of 50,000, or 750 per year.

Measuring adverse outcomes during pregnancy is often complicated by the fact that the exact time of conception is generally unknown. When the pregnancy is planned, the time from when a pregnancy is planned to a recognized pregnancy will be measured with reasonable accuracy. Estimating the rate of miscarriage requires registration of time from conception to the fetal loss in question or to the gestational week at which a fetal loss is defined as a stillbirth. At best, the time of conception may be identified by means of biochemical measures at a very early stage, but, in most cases, a diagnosis of pregnancy is not established until 3 to 4 weeks after conception at the earliest, and then it is retrospectively estimated by means of last menstrual period data (LMP) or – later – by using growth measures based upon ultrasound examination. Observation time for calculating rates of miscarriages, therefore, often starts at different time periods in gestation, and we have to take this delayed entry into consideration in order to obtain meaningful results when we try to, e.g., identify determinants of miscarriage (Baird et al. 1993).

45.5.1 The Measures Used to Describe Mortality

The World Health Organization (WHO) defines live births, fetal deaths, and induced abortions in the following way:

The definition of a *live birth* is the complete expulsion or extraction from the mother of a product of human conception, irrespective of the duration of pregnancy which, after such expulsion or extraction, breathes or shows any other evidence of life, such as beating of the heart, pulsation of the umbilical cord, or definite movement of voluntary muscles whether or not the umbilical cord has been cut or the placenta is attached.

Fetal death is defined as death prior to the complete expulsion or extraction from the mother of a product of human conception, fetus and placenta, irrespective of the duration of pregnancy: the death is indicated by the fact that, after such expulsion or extraction, the fetus does not breathe or show any other evidence of life, such as beating of the heart, pulsation of the umbilical cord, or definite movement of voluntary muscles. Heartbeats are to be distinguished from transient cardiac contractions; respiration is to be distinguished from fleeting respiratory efforts or gasps.

This definition excludes induced terminations of pregnancy.

Induced termination of pregnancy is defined as the purposeful interruption of an intrauterine pregnancy with the intention other than to produce a liveborn infant and which does not result in a live birth. This definition excludes management of prolonged retention of products of conception following fetal death.

Induced Termination of Pregnancy Rate (Conceptions). This measure uses live births, induced terminations of pregnancy, and fetal deaths in the denominator.

$$\begin{array}{l}\text{Induced termination}\\ \text{of pregnancy rate}\\ \text{(conceptions)}\end{array} = \frac{\begin{array}{l}\text{Number of induced terminations occurring during}\\ \text{a specific time period}\end{array}}{\begin{array}{l}\text{Number of induced terminations} + \text{live births} + \text{reported}\\ \text{fetal deaths during the same time period}\end{array}} * 1{,}000$$

Induced Termination of Pregnancy Rate (Population). This is the probability that women of reproductive age will have an induced termination of pregnancy within a given time period.

$$\begin{array}{l}\text{Induced termination}\\ \text{of pregnancy rate}\\ \text{(population)}\end{array} = \frac{\begin{array}{l}\text{Number of induced terminations occurring during}\\ \text{a specific time period}\end{array}}{\text{Female population aged 15 through 44 years}} * 1{,}000$$

$$\text{Fetal death rate} = \frac{\text{Number of fetal deaths during a specified time period}}{\begin{array}{l}\text{Number of fetal deaths + number of live births during}\\ \text{the same time period}\end{array}} * 1{,}000$$

$$\text{Fetal death ratio} = \frac{\text{Number of fetal deaths during a specified time period}}{\text{Number of live births during the same time period}} * 1{,}000$$

Maternal mortality ratio is the number of deaths attributed to maternal conditions in a given time period divided with a number of live births during the same time period.

WHO recommends including maternal deaths that occur within 42 days of the end of the pregnancy. Some countries use other time periods (i.e., within 1 year).

Although international comparisons are difficult to make because of variable reporting practices, we know that wide differences in maternal mortality exist worldwide (AbouZahr 1998).

The *perinatal mortality ratio* is the number of fetal deaths (>24 weeks of gestation) and deaths during the first 7 days of life divided by the number of stillbirths and liveborn children in the same period. Stillbirths are births of fetuses that show no sign of life after 24 weeks of gestation (or other specified thresholds).

Recent results indicate that this definition should be separated into stillbirths and deaths during the first week of life. Stillbirths should furthermore be divided into death before labor and death during labor. In the past, fetal death and early death after birth often had asphyxia as the common cause. Congenital malformations are now a much more common cause of death around the time of birth (Kramer et al. 2002).

Infant mortality is computed as the number of deaths during the first year of life, divided by the number of live births during the same time period. Instant mortality rates vary between 5‰ and 10% in the poorest countries in the world.

Many of these measures are difficult to record in a comparable way over time and between countries. Live births are well registered in many countries, but that is not the case with stillbirths, in part because the gestational age threshold that separates miscarriages from births differs between countries and in part because stillbirths do not count in population statistics (Gourbin and Masuy-Stroobant 1995).

In some countries, the threshold of 24 weeks is used to distinguish a birth from a miscarriage. Other cut-offs have been 28 or 27 weeks. Some countries use 20 weeks as the threshold. The complicating issue is, however, that birth of a liveborn child is a birth, regardless of the time of delivery or if the child dies shortly after birth. Any researcher with an interest in both live and stillbirths who uses routine registration systems to identify births should make sure that the time at risk for the outcome of interest is comparable (e.g., including in the analysis only babies born from the 28th week onwards, if that is the threshold for defining a stillbirth). It is often very difficult to identify the cause of death for stillbirths (Winbo et al. 1997) or the time of death, which may be of importance in a monitoring system. It has, for example, been suggested that fetal death during labor is a better indicator of the quality of obstetric care than fetal death before labor (Kiely et al. 1985).

45.6 Fetal and Infant Death

When we study mortality, we usually estimate mortality rates, i.e., the number of people who die as a function of the size of the underlying population and the period of time during which this population was under observation. Since mortality rates strongly depend upon age (and sex), we usually calculate age (and sex)-specific mortality rates. The mortality rate for people of 90 years of age will be the number of 90-year-olds who die within a year divided by the number of observation years we have for 90-year-olds in that population. We will count observation time from their 90th birthday until they turn 91, die, or leave the population for other reasons.

In reproductive epidemiology, age is even more important, but the problem is that we now work on two time schedules: one is starting at the time of conception and the other at the time of birth. Age may be counted from either the time of conception (gestational age) or the time from birth (normal age). We expect mortality to be high shortly after conception, and we also expect mortality to be high when the fetus leaves the intrauterine environment. We would prefer to present mortality as a function of observation time in the population at risk from conception, which is difficult for early fetal deaths (miscarriages) because we usually do not know how many have conceived in the source population. We have to start our observation when the pregnancy is diagnosed, and that often varies. Sometimes the pregnancy is first diagnosed by the occurrence of a miscarriage, and sometimes a conception ends in an early loss without the woman ever realizing that she was pregnant. Moreover, bleeding early in pregnancy may cause problems in the attribution of gestational age based on the date of last menstrual period (LMP) (Gjessing et al. 1999).

An additional problem is that the timing of fetal death is often not known and only the time of expulsion is. In early gestation, these two points in time may differ by several days or even weeks (missed abortions), and some dead fetuses may even be absorbed rather than expelled. Later in gestation, we expect a stillbirth to be closer to the time of death. When signs of life disappear, most women in affluent societies will seek medical assistance. The fetal death will be diagnosed and a birth induced.

Some studies on fetal death use survival methods that take delayed entry and gestational age into consideration, or they use the ratio of all miscarriages to births as the endpoint. In any case, caution is called for. If exposures cause very early (preclinical) loss, these need not be detected. Exposures that move miscarriages to the preclinical stage will appear as if they prevented miscarriage, regardless of whether they are based upon rates or cumulative risk.

The time in which a pregnancy is diagnosed depends upon a number of known and unknown factors. It is reasonable to expect a planned pregnancy to be detected earlier than an unplanned one. On average, women with regular menstrual cycles probably become aware of a pregnancy earlier than women with irregular cycles. A woman who has been pregnant before may detect symptoms of a pregnancy earlier than a woman who is pregnant for the first time. Availability and sensitivity of pregnancy tests will also play a role for the starting point of observation.

Assume that we base our study upon a cohort of pregnant women. Assume furthermore that we take an interest in an exposure that, for some reason, correlates with how early the pregnancy is recognized. The exposed women would then enter the cohort at a later (or earlier) gestational age than the unexposed women. Since the risk of miscarriage decreases with gestational age, at least early in pregnancy, their ratio of spontaneous abortions to birth would be lower (or higher) than that of the unexposed group and the results would thus be biased. Comparison of gestational age-specific miscarriage rates should, however, be unbiased – provided that the exposure does not modify the time period from fetal death to expulsion.

If an exposure changes only the timing of pregnancy detection, gestational week-specific ratios of miscarriage remain valid. If the exposure changes the timing of miscarriage around the threshold of when a loss becomes clinically detectable, all the possible measures of miscarriage may be biased.

Studying miscarriage may also be complicated for other reasons. Since the risk varies largely with gestational age, it is important to use valid data for gestational age in the model that is left truncated at the time of entry. Unfortunately, precise data on gestational ages are difficult to obtain, even if ultrasound measures are available. A fetus with a poor survival prognosis may show early growth patterns that deviate from standard values, possibly resulting in biased ultrasound estimates. Furthermore, the exposure of interest may also correlate with the estimate of gestational age (Henriksen et al. 1995), which may make it impossible to obtain unbiased comparisons even when doing the proper statistical analyses. In addition, the start of observation time need not coincide with the time of exposure. If the exposure causes miscarriage after a short time, the susceptible pregnancies may be removed from the study early on (and before they are recruited in a study), leaving only a selected group available for study. This selection could attenuate or even eliminate an effect of the exposure on the risk of miscarriage.

Using a case-control approach to study causes of miscarriage may be prone to bias in some situations. Controls should (using the principles of incidence density sampling, see chapter ▶Case-Control Studies of this handbook) be selected at the time of fetal death (which is often unknown) and not at the time of miscarriage. Furthermore, measurements taken at the time of miscarriage that change over gestational time may be poor indicators of the cause of fetal death, even in situations where they are the cause of death rather than a consequence of it.

Dietary habits (such as coffee consumption during pregnancy) may change when nausea disappears, and since a fetal death would reduce nausea, cases may then have a higher intake of coffee than controls, not because coffee killed the fetus but because nausea and aversion against coffee disappeared when the fetus died. The exposure frequency is then high during the time from death to expulsion but was not so at the time of death. The high coffee intake is thus a consequence of fetal death, not a cause of death, which is an example of reverse causation. Whether coffee consumption during pregnancy is associated with an increased risk of miscarriage continues to yield conflicting findings, even when trying to address the above issue (Weng et al. 2008; Savitz et al. 2008).

At birth, a new time schedule starts. Preterm or very preterm children will start this new clock before their fetal maturation has come to its natural end. Babies born in week 35 will start their extrauterine life 5 weeks before a baby born at term. Diseases that originated in utero with a fixed induction time will have an onset that perhaps should have been counted from conception time rather than the time of birth. Childhood colic may, for example, be a disease that peaks at a given time in childhood counting rather from the time of conception, independently of the time of birth (Sondergaard et al. 2000).

Measures that use births as the denominator rather than the population at risk in the proper time intervals deviate from the principles on which normal age-specific rates usually rest upon. Fetal death rates would, in normal practice, be seen as death within a given gestational time period divided by the observation time for fetuses at that gestational age, just as infant mortality is estimated as death during the first year of life divided by observation time for children less than 1 year of age in that population.

Some researchers estimate the week-specific rate of stillbirth (usually starting at week 20 or later). Conventionally, this has been done by dividing the number of stillbirths at each week by the number of stillbirths + live births in the denominator. This approach will result in the proportion of stillbirths decreasing with increasing gestational week since live births increase with gestational time.

Yudkin et al. (1987) have argued that it is incorrect to use all births at any given week to estimate the rate of week-specific antenatal fetal death rate. Instead, they suggested the so-called “fetuses at risk” approach, where all fetuses at risk of stillbirth are included in the denominator at each week (thus including all as yet unborn fetuses). Following this approach, which is more properly a rate than the conventional method, will yield a completely different impression of the stillbirth risk compared with the conventional method, i.e., the rate of stillbirth will increase at each gestational week, instead of decreasing. Joseph (2004) has proposed to extend this approach to neonatal deaths as well as other conditions. However, using this approach for postnatal events seems problematic, as it assumes that birth has not affected the babies’ risk (Paneth 2008). Furthermore, stillbirths are not at risk of neonatal death. Whether it is more appropriate to use the conventional or the fetuses at risk approach probably depends on the purpose of the analysis.

Rather than using survival principles in studying determinants of miscarriage, many use the ratio of miscarriages to birth, or miscarriages/(induced abortions + miscarriages + births). The latter presents a proportion of miscarriages among all who end their pregnancy with either a birth or an abortion. The first measure will overestimate the cumulative risk of miscarriage, as it does not take induced abortions into consideration (some of which would have ended in a miscarriage). The latter estimate will attenuate the cumulative risk since some of the miscarriages would have contributed to the numerator had they not been terminated. The frequency and the timing of induced abortion thus becomes a source of bias in these studies (Olsen 1984), unless data are analyzed by means of survival techniques. When examining a predictor of miscarriage that is also a strong predictor of induced abortion, such as age, it is particularly important to take terminations of pregnancy into consideration.

The risk that a pregnancy will end as a recognized miscarriage is high, especially if the mother is more than 35 years of age (Nybo Andersen et al. 2000). How high miscarriage rates are from the time of conception is not known because we only have data on conception for very specific in vitro fertilization (IVF) conceptions, which do not represent the general population. Kline et al. (1989) estimated that 50% of all conceptions end as a miscarriage, 40% of pregnancies that could be detected by hCG measures and 10% to 15% among clinically recognized pregnancies. These figures probably do not need much adjustment today, although a new cohort study among pregnancy planners found a slightly lower rate in the preclinical, but detectable, phase of pregnancy of about 25% to 30% (Hjollund et al. 2000), close to what Wilcox found in his study of pregnancy planners (Wilcox et al. 1988). No biological measure is at present able to pick up very early embryonic life (before implantation).

Many miscarried fetuses have chromosomal aberrations, especially among the very early losses (Macklon et al. 2002) as well as other congenital malformations (Shepard et al. 1989). Chromosomal aberrations, or more specific genetic defects, may be used to perform more detailed analyses of cause-specific fetal mortality. If all miscarriages were grouped together, the measure would represent a general mortality endpoint, and, since most exposures are expected to be specific in their causal action, a general mortality measure may in many situations be too imprecise for meaningful research. The problem is, however, that obtaining and genotyping fetal tissue when a miscarriage occurs is neither easy nor inexpensive in an epidemiological study that often includes large numbers of participants.

When studying a specific exposure that is not believed to cause chromosomal aberrations, it would be wise to restrict the outcome to miscarriages without such defects, if possible. Chromosomal analysis may, in some situations, be used to distinguish between consequences of fetal death from causes of the death itself. If, e.g., coffee intake is a result of fetal death rather than the cause of it, one should expect to see coffee equally associated with miscarriage with and without chromosomal aberrations. If coffee drinking is causally related to the fetal death, it will probably operate either via chromosomal aberrations or through another fetotoxic mechanism independently of the chromosomal aberrations. These analytical principles were to some extent used in a study by Cnattingius et al. (2000).

45.6.1 Birth Weight and Mortality

Reducing fetal and early childhood deaths is the goal of many public health programs. The best way to achieve this would be to remove or reduce the frequency of the underlying causes, especially if these causes also lead to long-term health problems for the babies who survive. Still, the main investments in most affluent societies have been spent on improving treatment, an approach due in large part to the fact that we do not know many of these causes or how to eliminate them. In some situations, effective treatment will not only save the life of the fetus but also lead to a better potential for a normal and healthy life. In other situations, treatment may increase the incidence of some conditions, for example, by keeping alive babies (often very preterm) who may be severely impaired.

A number of mortality measures, especially perinatal mortality rates, have been used to monitor how part of the health-care system performs. Perinatal mortality has declined over time in most countries and has reached low values in many industrialized countries and also in some less developed ones.

Often, mortality is presented stratified by birth weight in order to obtain a better monitoring instrument for the quality of treatment. The idea is that advanced treatment would especially show its effect on children born with a low birth weight (who are, generally, also born preterm), as these are the babies with the highest risk of not surviving early extra uterine life. However, birth-weight-specific mortality is fraught with problems and should be used with caution.

Wilcox and Russell (1986) showed that the strongest predictor of perinatal mortality in a population is the proportion of newborns with a birth weight that falls outside the population-specific Gaussian birth weight distribution (the residual). Birth weight usually follows a normal distribution with a small "bump" in the left side tail, mainly made up by small preterm births. The size of the residual (usually reported in percent of the total) is a more consistent predictor of the perinatal mortality for the population than the proportion of newborns with a low birth weight (<2,500 grams), which remains the main indicator used in monitoring systems. If we follow this thinking, the aim should not be that of changing the birth weight distribution for the population but to reduce the residual portion of the distribution, i.e., the proportion falling outside the Gaussian distribution. This translates mainly into preventing preterm birth rather than increasing birth weight. Most reproductive epidemiologists would probably agree on this strategy. The present obesity epidemic in many countries is, e.g., expected to increase birth weight in the coming years but not to decrease perinatal mortality; in fact, we may observe the opposite.

When stratifying mortality rates according to birth weight, a number of so-called paradoxes appear. For example, small babies born to mothers who smoke during pregnancy have lower mortality than small babies with low birth weights born to non-smoking mothers (among the first to comment on this apparent paradox were Yerushalmy (1964), Macmahon et al. (1965), and Comstock and Lundin (1967)). Babies born to smokers have, however, higher mortality at all birth weights beyond a certain point. The same is seen for African Americans compared with

European Americans in the USA (Adams et al. 1991; Wilcox 2001). Babies born at high altitude show a similar phenomenon, although their overall mortality is not elevated (Wilcox 2001). These phenomena of intersecting weight-specific mortality curves could be interpreted as a result of confounding by the underlying causes of impaired fetal growth. Smoking may, for example, be a less harmful determinant of low birth weight, in terms of mortality risk, than other causes that also cause low birth weight and mortality. Basso and Wilcox (2009) have shown how this apparent paradox could have a simple explanation (see below), without any causal link existing between birth weight and mortality. Heterogeneity in the mortality risk of different causes of low birth weight had been hypothesized as a mechanism for intersecting mortality curves by earlier publications (Macmahon et al. 1965; Meyer and Comstock 1972).

The nature of the relation between birth weight and mortality is a subject of debate (Chalmers 1979; Wilcox 2001; Shrimpton 2003; Haig 2003; Basso et al. 2006a; Hernandez-Diaz et al. 2006; Basso and Wilcox 2009). The main evidence for a causal link is limited to the strong association between the two; no direct evidence exists to causally link low birth weight and mortality. Basso et al. (2006a) have shown how, at least in theory, rare factors with a strong effect on both birth weight and mortality can reproduce the weight-specific mortality curve at a given week of gestation, without birth weight having any effect (direct or indirect) on mortality. The model is based on simple assumptions, illustrated in Fig. 45.1. Building on this premise, it can be shown that birth-weight-specific mortality curves will intersect when stratified on a factor, such as smoking, that has a weaker impact on mortality than the hypothesized confounders (Basso and Wilcox 2009), as shown in Fig. 45.2. According to this model, the reason why small babies born to smokers have lower mortality than equally small babies born to non-smokers is that there are worse reasons for being small than smoking (in terms of mortality risk). For a more detailed demonstration, see Basso and Wilcox (2009).

Although this explanation has been formulated assuming no effect of birth weight on mortality, such a strong assumption is not needed. However, for the intersection to occur in the specified manner, the "causal" association between birth weight and mortality would have to be considerably less (if any) than that suggested by the empirical weight-specific mortality curve (Basso and Wilcox 2009).

Stratifying perinatal mortality according to birth weight ranks (or a z-score) for the particular distribution rather than the absolute birth weight (e.g., calculating z separately for smokers and non-smokers) generally eliminates the paradox of crossing mortality curves, suggesting that birth weight in itself may be an inappropriate indicator of mortality risk. For the intersection to be eliminated through the z-score, however, it is necessary that the other factors that affect birth weight and mortality (other than the one being stratified on) be the same in the two groups (in terms of frequency and effect) (Basso and Wilcox 2009).

Birth weight has been widely used because it is available in most countries and generally well measured. Data on gestational age are, on the other hand, less often available, and the quality may be poor. Still, uncritical use of birth weight as an "exposure" or endpoint in reproductive epidemiology is not appropriate.

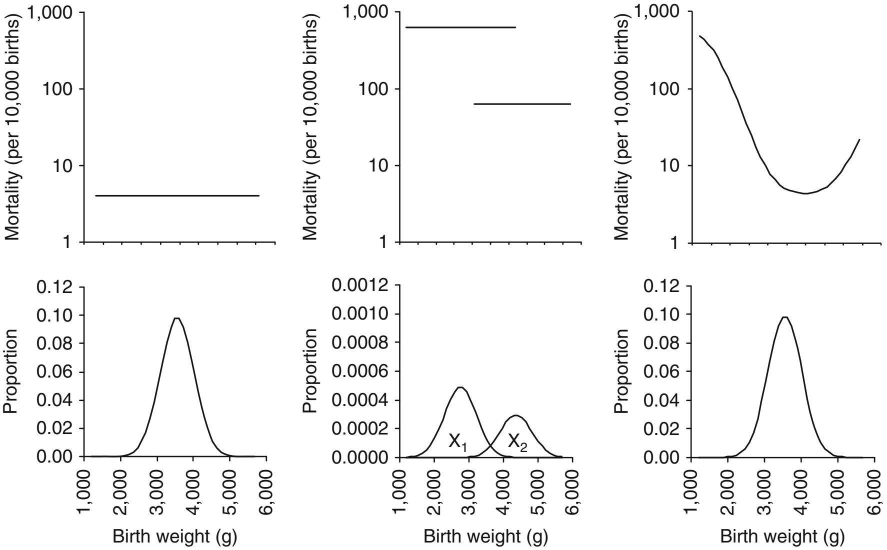


Fig. 45.1 Hypothetical scenario to reproduce the weight-specific neonatal mortality curve in the absence of any effect of birth weight on mortality. Population in figure ideally represents babies born at 40 weeks' gestation (Basso et al. 2006a; Basso and Wilcox 2009. Figure modified from Basso and Wilcox (2009)). *Left*: A population has mean birth weight of 3,500 g (±460) and mortality of 4 per 10,000 births. This distribution represents the birth weight that babies would have if no factor intervenes to modify their growth potential. *Center*: Two conditions, X_1 and X_2, affect 0.5% and 0.3% of babies, respectively. X_1 decreases birth weight by 782 g and increases mortality with an odds ratio (*OR*) of 171. X_2 increases birth weight by 782 g and increases mortality with an *OR* of 16 (note the different proportion scale in the y-axis of the birth weight distribution). *Right*: Resulting mortality curve plotted without stratifying by X_1 and X_2

If a woman is exposed to agents that reduce fetal growth, growth may be reduced proportionally or disproportionally. If mainly fat tissue is reduced, the newborn will have normal length but a reduced birth weight. The ponderal index is a measure that attempts to distinguish between thin and normal body proportions. The index is calculated as the newborn's weight divided by height raised to the third power. Readers, who are familiar with the body mass index (BMI), will know that the ponderal index deviates from BMI only by raising height to the power of 3 rather than to the power of 2. The only reason for this difference is to obtain a more symmetrical distribution of the ponderal index in newborns. How well this index actually reflects body composition among newborn children is, however, not well known, and its ability to predict adverse outcomes is also controversial.

There are other anthropometric measures of interest than birth weight or birth length. Head circumference is one such measure. Abdominal circumference may also be used. Most of these entities are probably measured with less precision, and less systematically, than birth weight, at least in countries that use standard weighing conditions and well-calibrated scales. It is also more difficult to accurately measure the length of a baby or a circumference. It is, e.g., likely that babies born vaginally

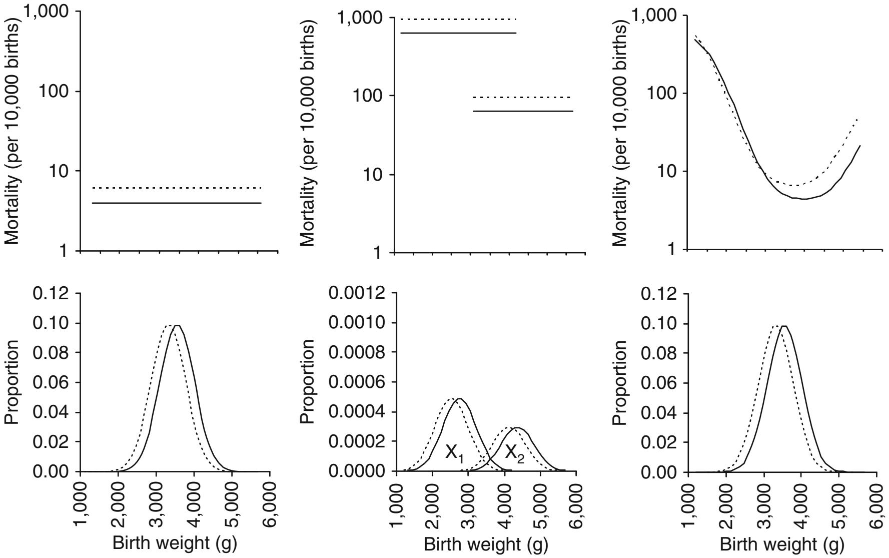


Fig. 45.2 A population of babies exposed to a factor F that decreases birth weight by 230 g (*broken line*) is added to the scenario shown in Fig. 45.1 (Figure modified from Basso and Wilcox (2009)). *Left*: Exposed population has standardized mean birth weight of 3,270 g (±460) and mortality 1.5 times that of non-exposed babies. *Center*: X_1 and X_2 affect 0.5% and 0.3%, respectively, of babies in each population. Effect of X_1 and X_2 on mortality of babies with F is increased with *OR*s of 171 and 16, respectively, (note the different proportion scale in the y-axis of the birth weight distribution). *Right*: Intersecting weight-specific mortality curves, stratified by exposure status

will present a molding of the cranial plates that will modify their head circumference compared to babies born with a caesarean section. Also, it is possible that some of these additional measurements would not be taken, which could result in selection bias if babies are excluded because not all measures were taken. It is also possible that some measures may have a systematically different quality depending on the baby's condition at birth (e.g., healthy babies may appear to be shorter than sick babies because it is more difficult to have them lie flat).

Since birth weight is a function of both pregnancy duration and fetal growth, pregnancy duration is usually taken into consideration when analyzing birth weight. The simplest procedure is to stratify results on preterm and term birth, but this will not fully account for the effect of gestational age on birth weight. Another popular option is to estimate "small for gestational age" (SGA), which implies identifying the, say, 10% (or 2.5%, or 5%) with the lowest birth weight among children born at each gestational week. Since the birth weight distribution is population-specific, an internal reference may be desirable if the size of the study allows for it. There are several disadvantages in using SGA; however, as it does not make use of the entire birth weight distribution, it will identify a number of babies who are constitutionally small but otherwise healthy, and it will consistently miss babies who have suffered from fetal growth disruption but have a birth weight above the designated threshold. Furthermore, at the earliest weeks of gestation, most births will be from complicated pregnancies, thus not representing the distribution of birth weight of all fetuses at that week (including those remaining in utero). This will result in the categorization of SGA defined on the basis of births to underestimate the true proportion of SGA babies when compared to the fetuses still in utero (Hediger et al. 1995; Hutcheon and Platt 2008). It has thus been proposed to use fetal standards, based on ultrasound measurements, although these will have a degree of measurement error (Cohen et al. 2010; Kacem et al. 2013). Babies who are defined as SGA in either manner do, however, have a substantially increased risk of death compared with those above this threshold.

SGA should, furthermore, not be taken as a measure of intrauterine growth restriction (IUGR) since the term SGA is purely descriptive. The use of the term IUGR should be limited to situations where it is known that the fetus is growth restricted. Such a diagnosis requires, at least in principle, a documented deviation from the growth curve for that baby, which may be available if several ultrasound measurements have been taken during pregnancy.

Gestational age, as the strongest predictor of birth weight, should be taken into consideration when continuous birth weight is the outcome. In order to account for the non-linear effect of gestational age on birth weight, gestational age could be modeled including a quadratic term, or with quadratic splines, or as several categories of gestational age as dummy variables (for an introduction to regression models, see chapter ▶Regression Methods for Epidemiological Analysis of this handbook).

The drawback of adjusting birth weight for gestational age one way or the other (or to use SGA measures) is mainly related to the often poor quality of data for gestational age, especially evident at some low gestational ages, where babies that are most likely born at a later gestation (given their birth weight) are instead recorded with a low gestational age (Gjessing et al. 1999; Tentoni et al. 2004; Basso and Wilcox 2010). Birth weight is usually measured more accurately than gestational age, and, by using an endpoint such as small for gestational age (SGA), the quality of the data may be compromised by making use of a composite measure that includes a variable that is imprecise at best and possibly biased (although gestational age is a very important variable). An exposure with no effect on fetal growth that causes irregularities in the menstrual cycle could either show an effect on SGA, if gestational age was based upon LMP data, or be biased if gestational age is measured by ultrasound and the exposure correlates with early fetal growth (Henriksen et al. 1995; Henriksen and Wilcox 1994).

The term “small for gestational age” is somewhat misleading, since it is a purely descriptive population concept: the baby is among the smallest in this particular population. Usually, what we would like to know is if the baby is small because it has suffered a disruption in growth. A baby that has achieved its full genetic growth potential could be an SGA baby just because it is constitutionally small. We expect such a baby to be at low risk for complications and diseases that may be related to impaired fetal growth.

Our interest in birth weight from a health perspective should focus upon a deviation from the genetic growth potential rather than on the absolute birth weight. Unfortunately, we do not know the genetic programming of the fetus, and we often have to rely upon indirect estimates that could lead to severe misclassification. The birth weight of babies previously born to the same mother could be used as a tool for estimating the growth potential of a later baby (Skjaerven et al. 2000; Basso et al. 2005; Pedersen et al. 2007), as the correlation between birth weights of siblings who share the same mother is about 0.5.

Given the fact that birth weight has been used for convenience – too extensively and for too long a period – a group of scientists (Adams et al. 2003) met in June 2002 in Denmark to announce the so-called Sostrup statement (named after the residency where the meeting took place). These statements concluded that:

1. In population studies, "percent low birth weight" (LBW) is a poor research tool for detecting factors or conditions that damage perinatal health.
2. "Percent LBW" is a poor index of population perinatal health.
3. Adjustment for absolute birth weight is rarely justifiable in looking for effects of specific exposures on infant or perinatal mortality.
4. Some exposures or conditions may compromise fetuses without causing preterm delivery or impaired fetal growth.

It is, however, difficult to propose another indicator that is as accurately measured and as easy to obtain, especially if we want such an indicator to be applicable in less affluent countries. The best alternative for a single indicator is probably preterm birth, although that is subject to a substantially higher degree of misclassification, especially when estimated on the basis of the last menstrual date.

45.6.2 Optimal Birth Weight

The concept of "optimal birth weight" has been used in the literature mainly to indicate the birth weight with the lowest perinatal mortality in specific populations. It is, however, unclear whether this birth weight actually is associated with lower mortality because of some biological property or, instead, because it is the category of birth weight associated with the "optimal" combination of factors that affect birth weight and mortality. As Basso and coauthors show (Basso et al. 2006a; Basso and Wilcox 2009), a model that assumes no effect of birth weight will nonetheless result in an "optimal" birth weight slightly above the mean. This happens because the factors that reduce birth weight and increase mortality have a greater impact than those that increase birth weight and mortality, a possibility previously suggested by Haig (2003). Whether or not birth weight is believed to play a role, it is probably legitimate to assume that babies born at the optimal birth weight are the least likely to have suffered from severe fetal growth disruptions, although no interpretation of the causal effect of birth weight on mortality can be currently ascribed to this weight.

Overall, we do not recommend the concept of optimal birth weight to be used broadly, as the optimal birth weight depends on what the birth weight has to "optimize," mortality, immune defenses, cognitive development, etc. In general, it is to be expected that the "optimal birth weight" would be a different birth weight for any given baby and that it would follow an approximately normal distribution at each week of gestation (as represented by the population in the left panel of Fig. 45.1).

The problems related to the concept of "optimal birth weight" without taking into consideration what its determinants are illustrated by the importance of the mother's BMI on birth weight (Nohr et al. 2008). The obesity epidemic will increase birth weights in the population but without reducing morbidity and mortality.

45.7 Gestational Age

A pregnancy starts at conception, but, since the time of conception is usually unknown, the starting point is often taken from the first day of the last menstrual period (LMP), which is usually around 2 weeks prior to conception. Based on this, a pregnancy is expected to last 280 days, or 40 weeks.

In most countries, Naegele's rule is still applied to estimate gestational age when using LMP data. The expected day of birth is calculated starting from the first day of LMP, and then 3 months are subtracted and 1 week is added. This rule, however, works well only for women who remember their menstrual periods and whose periods are regular (and close to 28 days of duration). The Naegele rule also works best in non-leap years at the population level (Basso et al. 2000a). With such easy access to electronic calendars, one would expect to see more electronic devices that simply add 280 days to LMP, taking leap years into account.

In affluent societies, ultrasounds (US) are more and more often used as a way of estimating gestational age, even in normal pregnancies with a certain LMP date. The idea behind the estimate is that certain structures, like the biparietal diameter (BPD), grow linearly and similarly for all babies in early pregnancy.

A given diameter is compared with a growth standard, and the gestational week is based upon the measure read from the standard growth curve – 16 weeks in the example above (cf. Fig. 45.3).

Experience has shown that ultrasound estimates are more precise than LMP measures and, for this reason, are better at predicting the date of delivery. They need not always be better for research, however. If you study exposures that impair early fetal growth, then an ultrasound measure may cause bias. The bias is probably too small to be of relevance in clinical practice, but it may be of concern in research. There may be research projects that are better off with an unbiased estimate with a low precision (like LMP) than with a biased estimate with a high precision. As previously mentioned, however, a substantial proportion of misclassified gestations is present in LMP estimates at pre- and postterm weeks. Thus, their use for research is unadvisable, especially when focusing on the extremes of the distribution of gestational length (Joseph et al. 2007; Basso and Wilcox 2010).

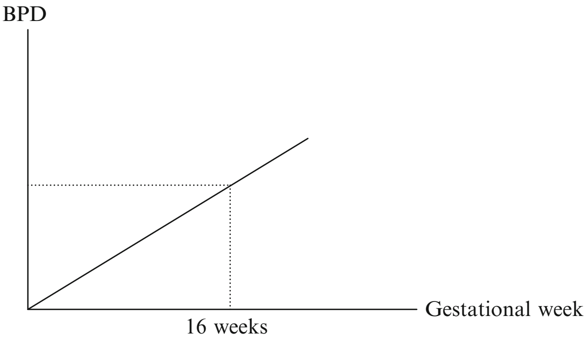


Fig. 45.3 Biparietal diameter (BPD) according to time since conception

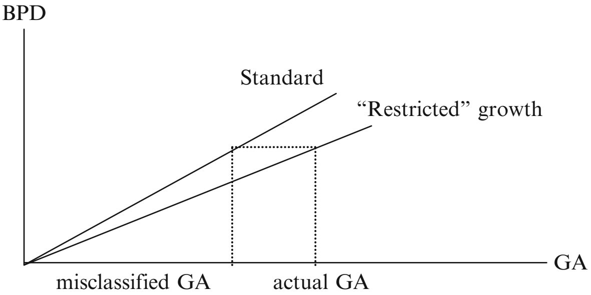


Fig. 45.4 Actual and standard Biparietal diameter (BPD) growth curves. GA = gestational age

As certain exposures, such as smoking, may affect even early fetal growth (Henriksen et al. 1995), use of ultrasound dating could result in babies of smokers to systematically be misclassified as having an earlier gestation than their actual one, given that their growth curve will be lower than the standard (Fig. 45.4). Assume that the difference between the actual and estimated gestational age (GA) is 2 days. If the woman gives birth shortly after 37 weeks of gestation, she would – erroneously – be classified as having given birth preterm.

If an inappropriate standard is used indicating a more rapid growth than for the population, gestational age will then be biased towards a low value, which will lead to a higher frequency of preterm birth. Using an inappropriate standard for the population under study will thus have impact on the estimated proportions of preterm and postterm birth for the population.

Gestational age is counted in days or in completed weeks. Preterm birth is a birth occurring before the woman has reached week 37, while postterm birth is defined as a birth taking place after completion of the 42nd week. "Prematurity" was often used in the past for newborns with a birth weight of less than 2,500 g. This term should generally be avoided, because we do not know if newborns with a birth weight of less than 2,500 g are premature; some will not be. Being premature requires clinical signs of prematurity; it is a clinical diagnosis that should not be used unless clinical data support it. "Preterm" is a more accurate description when having data on gestational age only, and sometimes it will be useful to study very preterm as well. Very preterm has been defined as birth before 34, 33, or 32 weeks

(Berkowitz and Papiernik 1993). Some studies investigate babies born even before 32 weeks. Thus, the definition of “very preterm” is generally study specific. The best definition may depend not only on the hypothesis one wants to examine but also on the available sample size, as very early births are, fortunately, very rare.

The frequency of pre- and postterm births depends upon precision and validity of the estimates of gestational age. If the central tendency of two measures (ultrasound and LMP) are the same, but one is measured with larger measurement errors (like LMP), then the imprecise measure will lead to more pre- and postterm births, provided that the central tendency is the same. In a study of pre- or postterm delivery, it is therefore important that gestational age is at least measured using the same methods in the groups to be compared. If studies are based upon routine registrations of gestational age, it may not be known how it was measured. If the exposure under study correlates with early fetal growth, LMP should be applied. If the exposure correlates with menstrual irregularities, then ultrasound is preferable. In order to detect a given difference in gestational age between two compared groups, a larger sample size is usually needed if gestational age is estimated by LMP compared with ultrasound estimates.

An example of the latter case may occur when studying whether a long waiting time to pregnancy leads to preterm delivery. If only LMP measures are available, one may find that women with irregular or long menstrual periods have a longer TTP, as many subfecund women have irregular cycles. The measure of time to pregnancy itself will be affected by this irregularity, since time to pregnancy, although ideally reflecting the number of cycles that a couple takes to achieve a clinically detectable pregnancy, is often reported in months. In such cases, not only imprecision but also bias can affect the effect measures, and the burden to estimate how much the observed effect can be ascribed to these problems falls on the researcher. If ultrasound measurements are not available, one way to assess whether the association may be due to bias caused by differential misclassification of gestational age could be to use birth weight as a support measure to corroborate the finding. Birth weight is less subject to error than gestational duration, and its accuracy is less, if at all, dependent on correlates of the exposure. However, birth weight is strongly influenced by duration of gestation. Thus, checking whether the association between a given exposure and preterm birth is replicated when using birth weight as the outcome may provide reassurance for the finding – or grounds for rejecting it. If there is no difference in the birth weight distribution according to time to pregnancy before adjustment for gestational age, this suggests that most of, or all, the observed effect is due to differential measurement error of gestational age. If, on the other hand, the birth weight of children born to women with a long time to pregnancy is lower than that of women with a short time to pregnancy, this is supportive of the conclusion that there may be some effect of time to pregnancy on gestational duration (provided that there are no residual confounders that can reduce birth weight that are more common in women with long time to pregnancy): this would be a problem if, for instance, short women – who have lighter babies – also had a longer time to pregnancy than tall women but a gestation of the same duration. A long time to pregnancy appears to have an effect on fetal growth as well as on gestational age (Zhu et al. 2007a; Basso and Baird 2003), and one could thus estimate this by adjusting for gestational age when examining birth weight as a function of time to pregnancy.

Another example of a factor affecting the accuracy of gestational age selectively in exposed and non-exposed women may occur when examining the effect of short interpregnancy intervals (the time between the previous birth and the next conception) on the gestational age of the subsequent pregnancy. Women who have short intervals are likely to have less accurate estimate of gestational age, because cycles resume some time after a pregnancy, and they may be irregular at first. Therefore, women with very short intervals may have systematically inaccurate measures of gestational age. Again, by checking with birth weight, one can try to assess whether there is evidence of this phenomenon. Although short interpregnancy intervals have been associated with both preterm delivery and low birth weight, in a study among Danish women (Basso et al. 1998), only the association with preterm delivery was observed, while the one with low birth weight disappeared entirely after adjustment for preterm delivery.

As previously stated, being born pre- or postterm may have stronger health impacts than deviating from a “normal” birth weight. Using the proportion of preterm births as an indicator of one component of reproductive health in a monitoring program is therefore of interest. In the past, such a measure was based upon LMP data. Now it would often be mainly ultrasound based. From a monitoring point of view, this raises issues related to comparability over time. The standard used to estimate gestational age based upon BPD should be appropriate, and that means both time- and population-specific. It should be a standard for the population it is applied to, and it should be changed over time if fetal growth in the population changes over time. In diverse populations, one standard may not be enough, if growth patterns vary by ethnicity. Many countries face increasing obesity problems, which may influence not only birth weights but also early fetal growth. If so, BPD standards need frequent adjustments for an ultrasound-based monitoring system of preterm births to be unbiased.

No matter how gestational ages are calculated, you will often find implausible gestational ages based upon the newborn children’s birth weight or maturity. Depending on the circumstances, one can attempt to “clean” the data or code such instances are missing. Whichever approach is taken, however, there will be consequences for the analysis. For estimating gestational age at the time of miscarriage, the LMP method is to be recommended in most cases, even if ultrasound was used.

It is well accepted that preterm birth is a heterogeneous condition (Romero et al. 1994, 2006; Klebanoff and Shiono 1995; Klebanoff and Schoendorf 2004; Greenwood et al. 2005; Savitz 2008; McElrath et al. 2008). Although being born early is a strong predictor of morbidity and mortality per se, the various causes of preterm birth likely have an independent effect on mortality risk. Thus, the observed week-specific mortality rates do not represent the pure effect of being born early but are a combination of both components. It is thus theoretically possible that a large portion of mortality among preterm babies is due to the pathologies that cause preterm birth and increase mortality risk (Basso and Wilcox 2010, 2011). Just as in the case of birth weight, remembering that gestational length does not explain the entire gradient of risk from very preterm to postterm can explain what may appear at first to be paradoxes, such as those instances where high-risk babies (e.g., twins, babies with pre-eclampsia) have lower mortality at preterm weeks than babies at lower risk (e.g., singletons, babies without pre-eclampsia). In other words, a twin baby born at 32 weeks is more likely than a singleton to have a more benign reason for being born preterm (as the only problem may have been twinning), whereas a singleton will have to have had another reason, potentially conferring a higher mortality risk (Basso and Wilcox 2010, 2011). Because babies born preterm tend to constitute a pathological group, studies among very preterm births (such as those carried out within perinatal networks) that examine morbidity or mortality as a function of a specific condition (e.g., pre-eclampsia and retinopathy of prematurity, see, e.g., Fortes Filho et al. 2011) cannot be interpreted causally, as babies with the exposure of interest are compared with babies who have other, often non-specified, exposures that can affect the outcome under study (Basso and Wilcox 2011). In general, as long as there are unmeasured causes of preterm birth that strongly affect the outcome, it is not possible to estimate the "pure" mortality (or morbidity) risk due to being born early, nor the causal effect of a given cause of preterm birth on an endpoint (Mann et al. 2011).

The pregnancy period is, in many countries, under intense monitoring by health-care personnel. The effects of any exposure under study reflect only the effect remaining after health-care intervention. This limitation is always true but is especially important to keep in mind when studying pregnancy duration, especially postterm birth. A birth may be induced when the clinicians believe the child is better off outside the uterus, or perhaps just because they believe the child is sufficiently mature to be born, or if the mother's health is at risk if the pregnancy is continued. In any case, a substantial number of pregnancies are not carried to a "natural" end, and these observations are "censored." Since the censoring will, in many cases, be associated with the risk of complications (to the mother or the baby), we cannot study, e.g., the risk associated with postterm birth per se. The only option is to study what remains of risk after health-care intervention. In like manner, one cannot nowadays study determinants of postterm delivery, only determinants for pregnancies that are allowed to continue after 42 weeks of gestation. Although these limitations are self-evident, they are often not mentioned in scientific reports – perhaps because they are self-evident (or are "forgotten"). In any case, the necessary precautionary warnings should be mentioned so that the results are not misleading.

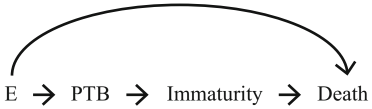
It is well known that many adverse outcomes strongly correlate with preterm birth (PTB) and/or low birth weight (LBW), although these associations may not be causal. These factors are probably the strongest predictors of early postnatal mortality and morbidity. PTB and LBW are, however, not isolated events but have causes that may directly or indirectly cause the reproductive failure of interest. LBW is, for example, strongly correlated with infant mortality, but the current obesity may well reduce the frequency of LBW without reducing mortality.

A failure to distinguish between a confounder and an intermediate may also have serious consequences as illustrated by a few simple directed acyclic graphs (DAGs) (cf. chapter ▸Directed Acyclic Graphs of this handbook).

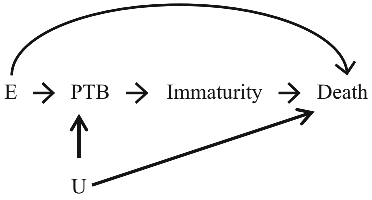
In the following example, an exposure's (E) effect on mortality is mediated by PTB and immaturity:

E → PTB → Immaturity → Death

Adjusting for these factors would block the association you want to study. But there could be a direct link from the exposure (E) to the disease (D) you take an interest in. If so the DAG looks like this:



The direct link from E → Death would show if you adjusted for PTB (or immaturity). Most likely the DAG would be a bit more complicated because there are other factors (U) causing PTB and the DAG could look like this:



If you now adjust for PTB, you will generate a backdoor path E – U – Death. This backdoor path could be closed by controlling for U. However, U usually stands for "unknown" or "unmeasured," leaving this option out.

Simply adjusting for length of gestation in the situation above may yield grossly biased results (Wilcox et al. 2011). Alternatively, assumptions could be made on the frequency and strength of the U's to obtain a range of estimates of the direct effect of E on the outcome, be it neonatal death or another endpoint (VanderWeele and Hernández-Diaz 2011; VanderWeele et al. 2012; MacLehose and Kaufman 2012).

A failure to distinguish between confounders and intermediates has produced many misleading results that need to be corrected. The problem with intermediating factors has been well known for many decades, although this has not prevented misleading analyses. The awareness of collider bias is, however, relatively new and a result of using DAGs in causal research. DAGs have also been helpful in our understanding on how to deal with past pregnancy histories (Howards et al. 2012).

45.8 Congenital Malformations

Congenital malformations have been subject to numerous epidemiological studies and monitoring systems because they are very serious for the affected children, their families, and society in general. Researchers dealing with studies of teratogenic or fetotoxic effects prefer to include a wider range of abnormalities than structural malformations. Some prefer to use the broader term congenital anomalies (CA) that will include genetic disorders and some functional impairment as well. CA will be present in 2% to 7% of all newborn children, depending on the definition and the level of monitoring. Many anomalies will not be diagnosed until childhood or even later, and a number of defects (such as some heart malformations) may go undetected for many years or not ever be diagnosed. The effectiveness of prenatal screening followed by induced abortion of affected fetuses also plays a role for the prevalence of congenital anomalies at birth.

If congenital anomalies are taken to be all structural or functional defects or deviations that are present at birth (whether or not they are detected at the time), the frequency could be defined to cover many more than 7%. Many functional defects may be present at birth in a form that is not yet detectable with the diagnostic tools that we have at present, such as cognitive defects, color blindness, other brain defects, mutations of importance for some cancers, like childhood leukemia, or testis cancer. Fetal organ programming of organ functioning could, in principle, also be seen as a congenital anomaly in the sense that programming may increase susceptibility to many diseases, such as insulin resistance and all the diseases it may lead to. Clearly, using such a broad approach makes congenital abnormalities an impracticable or even impossible endpoint. Most studies and monitoring systems will restrict the endpoint to what is described in official disease classifications, such as Chap. 17 in the International Classification of Diseases (ICD10).

Monitoring systems of congenital malformations have been, and are still being, used in order to detect changes in the prevalence of malformations over time. Many of these systems have their root in the Thalidomide disaster. Thalidomide was released on the market in the 1950s as a sedative and a tranquillizer to treat nausea and insomnia and was considered to be safe even for pregnant women. The drug was regularly used by pregnant women in countries where this drug was approved. The drug was teratogenic, as reported by Lenz and MacBride in 1961 (Diggle 2001). About 40% of the pregnant women who used the drug during organogenesis (mainly in the second and third months of gestation) gave birth to children with malformations. The most common defect was phocomelia, a syndrome where the extremities were severely underdeveloped. The drug is not teratogenic in most experimental animals, and pregnant women are never included in pre-marketing randomized trials. Pregnant women used the drug believing it was safe. It is therefore important to set up programs that fully utilize the information generated by pregnant women when they start using new drugs. At present, we are alarmingly short of such information systems (Olsen et al. 2002).

It is currently hoped, somewhat naïvely, that the monitoring systems will pick up new teratogenic drugs, even at an early phase, and that such an effect will be detected by the reporting of side effects. Most drugs, especially new ones, are, however, only used by few pregnant women, and the person who prescribes the drug is usually not the same as the person who diagnoses a congenital malformation. A monitoring system of congenital malformations would, on the other hand, be of importance for setting up specific studies, because good quality data on congenital malformations are often lacking in routine medical records.

The technology available to detect congenital malformations at an early stage in gestation is continuously improving with the use of ultrasound or biochemical and genetic methods. Clearly, measuring prevalence at birth may become a questionable endpoint if prenatal diagnosis is not used in exactly the same way in the groups to be compared.

Since specific malformations are rare, most studies on determinants of congenital malformations will be based upon large routinely collected data sources covering many thousands, possibly hundreds of thousands, of pregnant women, or by using a case-control approach. The main advantage with the case-control approach is the possibility of collecting valid exposure information concerning the pregnancy (often very early pregnancy) by means of interviews. These data may, however, be difficult or even impossible to obtain for exposures that are hard to recall. The recall easily leads to bias related to a lack in symmetry of the information obtained from a woman who had a child with a severe handicap when compared to the information provided by a woman who had a healthy child. Using another set of patient controls (e.g., another type of congenital malformation than the one under study) may be a possible solution, although it is not without problems. When using patient controls as a surrogate for representative source population samples, the "control" disease should neither be caused nor prevented by the exposure under study, and, since we know so little about the causes of most malformations, this might be a hazardous assumption to make. However, interviewing women who all had a child with congenital malformations should render the quality of information more comparable, thus reducing the potential for bias (Swan et al. 1992). There is indication of a more accurate recall of medicine intake in mothers who had a child with a congenital malformation than in mothers who had a healthy child (Rockenbauer et al. 2001). Furthermore, it has been shown that the way questions on medicine intake are phrased plays an important role (Mitchell et al. 1986).

Using a case-crossover design (Maclure 1991) might be an option to overcome biased reporting (and confounding by personal characteristics). Since only exposures at a given time period may be of relevance for the malformations under study, another time interval during gestation before or after the index exposure may be selected as a reference. The relative prevalence ratio for the case in question can be estimated by dividing mothers who were exposed in the exposure window but not in the reference window and vice versa, given that these windows have the same duration. The rest of the exposure combinations do not provide any information to the study concerning the associations between drug intake and the specific malformations (Kjaer et al. 2007, 2008).

A case-control approach will allow specific diagnostic classification of the cases. By setting up specific diagnostic standards, it is possible to make sure that only the individuals with the malformation in question enter as cases (high specificity), although some with the disease may not meet the criteria (sensitivity <1). As long as specificity is high, however, a relative effect measure will not be biased by a low sensitivity, but the study power will be reduced. The following example illustrates why this is the case.

Imagine that the underlying source population has the following structure:

Exp	Cases	All births
+	2,000	100,000
–	1,000	100,000

$$RP = \frac{2{,}000/100{,}000}{1{,}000/100{,}000} = 2.0$$

Now assume that only half of the time cases will be diagnosed. The display of data would then be:

Exp	Cases	All births
+	1,000	100,000
–	500	100,000

$$RP = \frac{1{,}000/100{,}000}{500/100{,}000} = 2.0$$

The relative prevalence ratio (*RP*) is still 2, simply because there are still twice as many diagnosed children with malformations among the exposed compared with the non-exposed. The low diagnostic sensitivity did not change this. It is easy to design a case-control study to replicate these results. It is just a matter of proper sampling from the source population.

A case-control sampling using all cases and a sample from the base from any of the two above situations would produce an unbiased estimate of the *RP*. A 1:5 case-control sampling would, in the first case, give

Exp	Cases	Controls
+	2,000	7,500
–	1,000	7,500
All	3,000	15,000

$$OR = \frac{2{,}000/1{,}000}{7{,}500/7{,}500} = 2.0$$

and in the second case

Exp	Cases	Controls
+	1,000	3,750
–	500	3,750
All	1,500	7,500

$$OR = \frac{1{,}000/500}{3{,}750/3{,}750} = 2.0$$

The power would be less in the second example, as reflected by a larger variance in the second study (the variance of the odds ratio (OR) is 0.0018 and 0.0035 in the first and second study, respectively).

45.9 Pregnancy Complications

45.9.1 Operational Definition

During pregnancy almost all maternal physiological systems are subjected to major changes. Cardiac output increases by 30–50%, and thus the kidneys have to filter a much higher amount of blood. The space taken up by the enlarging uterus changes the way the lungs and digestive systems work. The major changes are, however, hormonal, and the placenta produces a large amount of hormones that help maintain the pregnancy. The immune system is also affected, as pregnancy is a mildly immunodepressed state. Also, there are indications that pregnancy requires a shift in the type of immune response from Th1 (pro-inflammatory) to Th2 (antibody mediated), modifying the maternal type of immune response as well.

Because of these major changes, there are a number of preexisting diseases (such as diabetes, kidney disease, affections of the thyroid, heart failure, rheumatoid arthritis, and other autoimmune diseases) that may be modified by pregnancy and harm the woman or the fetus. Some autoimmune diseases improve in pregnancy, such as rheumatoid arthritis; others relapse or worsen or present a cluster of onsets immediately following a pregnancy.

Pregnancy complications are defined in the Medical Subject Headings as the co-occurrence of pregnancy and a disease. The disease may have started before conception or after. This definition is rather general and not in line with what most people would think of as being a pregnancy complication. A puerperal depression is considered triggered by the birth and not just a depression that happens to occur shortly after delivery, although it may be difficult to distinguish a depression triggered by birth from a depression that would have occurred in any case. In these paragraphs, we will, therefore, deal with some aspects of complications that are only seen during pregnancy or which are only defined as such during a pregnancy. We will briefly deal with some common pregnancy complications (such as hyperemesis and placenta previa), reserving a particular emphasis to gestational diabetes and pre-eclampsia, the former because it is a clear example of some of the problems facing epidemiologists who deal with pregnancy complications. Pre-eclampsia is a relatively common and serious complication of pregnancy, and it is also one of the most fascinating mysteries of reproduction.

Other chronic diseases pose a risk to the mother and to the fetus. It is likely that most of them will be seen among pregnant women, but with a lower prevalence than in the general population, especially if they are severely debilitating for the woman in her reproductive years. The "healthy pregnancy effect" is analogous to the better known "healthy worker effect" and is due to the fact that a reasonably good health is required to conceive and carry a pregnancy to term. This "effect" need not cause bias if properly addressed in the design of the study. The "healthy pregnancy effect" only underlines the fact that many diseases will be less frequent among parous women if these diseases interfere with actual fertility.

45.9.2 Methodological Challenges

45.9.2.1 Defining a Disease: Choosing a Cut-Point

A general problem in studying pregnancy complications is due to the fact that many such conditions are defined as an extreme of events that occur in the course of a normal pregnancy. Thus, a disease that represents an extreme value of the distribution of a given trait rather than a qualitatively different entity will be more problematic to study. In this case, the distinction between normal and pathological becomes relatively blurred, and very often the challenge faced by clinicians is that of defining a cut-point beyond which a condition is declared a disease (as in obesity, diabetes, hypertension, and pre-eclampsia) and below which the same condition is considered within the norm. It is immediately evident that definitions of this type are susceptible to many problems, as is the case with all diseases defined this way, because any arbitrary threshold will introduce some degree of misclassification, especially since pregnancy is a condition under intensive medical surveillance, which will then make pregnant women a population in which virtually the entirety of its members will be screened for severe diseases one way or the other and, in most cases, more than once in the course of pregnancy. It is well known that even a test with a high specificity will produce a large number of false positives in a population with a low prevalence of a disease, and this leads to women being unnecessarily treated and subjected to the stress of being diagnosed with a disease they may not have. However, missing women with a given disease by moving the cut-point towards more extreme values will result in a low sensitivity that will lead to missing cases with consequences that may be very serious for the mother and the baby. The extent of these problems depends not only on the criteria for defining a disease but also on the approach adopted by the care providers, the access to prenatal care, and the frequency with which pregnant women are monitored. The more frequently women are seen and their blood glucose, blood pressure, or proteinuria are measured, the more likely it will be that they can be wrongly (and rightly) classified as having a given disorder, especially if these measures fluctuate over time. Conversely, women who are not screened often or do not comply with prenatal care may be underdiagnosed in these circumstances, and women who are frequently screened may be overdiagnosed. Factors related to monitoring, such as insurance coverage and distance from antenatal care centers, can thus produce bias, especially – but not only – if risk factors for the disease are also part of the reasons why women comply less with prenatal care (smoking may, in some circumstances, fulfill these criteria). In some cases, modeling the number of ANC visits or other factors affecting the access to ANC (distance, social class, insurance plan when applicable) might provide some information about whether a problem of this type has occurred, but this will not necessarily be sufficient to correct for the bias.

Even if all women comply equally well with prenatal care, problems may arise if the exposures under study correlate to some degree with the probability of being diagnosed, as health-care personnel may differentially screen women according to their risk profiles.

Furthermore, the consensus on the cut-points usually changes in time and is often not even geographically homogeneous at the same point in time. Comparisons over time and between different areas thus become difficult to perform and of questionable value depending on the level of prenatal care and definition of the disorder. Often researchers do not have the crude values of what is actually measured but only the clinical diagnosis, which makes virtually every study susceptible to well-founded criticism because uncertainties and the potential inaccuracy of the diagnosis may well depend on the putative risk factors under study. Random errors will also tend to dilute associations with the number of false positives that will be included in most case series. We are thus left with studying the phenotype that clinicians in a particular region and at a given point in time call a disease, which may not be the best classification from an etiological point of view.

45.9.2.2 Design Strategies

In general, i.e., if the disease prevalence is low, relative effect measures, such as the relative risk or the odds ratio, are more profoundly affected by low specificity than by low sensitivity. A possible strategy when dealing with diseases that have been classified on the basis of a definition determined by exceeding a given cut-point (such as gestational diabetes) is to limit the number of false positives by restricting the study to the more severe cases (which will, most likely, be identified with a lesser degree of error), if the possibility to do so exists. However, this may reduce the power of the study when dealing with disorders that, in general, are quite rare to start with.

Ad hoc studies (made for the specific purpose of investigating a given disease in a defined population) with access to medical records are an important option, which, however, will be more costly than studies based on routine information recorded, for example, in registries and will, once again, raise the issue of how many cases will be available for the study.

In some instances, combining the clinical diagnosis of the disease of interest with a feature that should be concomitantly present if the disease is actually present may increase the quality of the data for the study. For example, placenta previa is a relatively rare pregnancy complication, where the placenta is positioned to partly cover the opening to the birth canal. In early pregnancy, however, the placenta will relatively often appear to be positioned in such a way but will, in many cases, spontaneously reposition itself during the course of pregnancy (Dashe et al. 2002). If a study is based on all diagnoses of placenta previa, including those made in early pregnancy, cases will include a proportion that are not truly cases. If placenta previa persists, on the other hand, the common practice is that of performing a caesarean section to deliver the baby and thus, by including only the women diagnosed with placenta previa who were later delivered by caesarean, one will probably be limiting the analysis to the more severe or, at least, the more persistent cases. This has been done, for example, in studies investigating whether pregnancies with placenta previa presented a higher male to female ratio (Wen et al. 2000; Ananth et al. 2001). Whether this type of approach should be used or not has to be evaluated depending on the specific aim of the study.

The significance of the same symptom may change across the duration of pregnancy, and it may be a cause or a consequence of another condition, also depending on the timing. This is probably the case, for example, of bleeding during pregnancy, which, in the last trimester, is probably a consequence of placenta previa or placental abruption, while it may be an entirely different entity in the first two trimesters (such as threatened abortion). Hyperemesis, severe vomiting in pregnancy, is another example of a "normal" condition becoming pathological when presenting with extreme symptoms. Several studies have investigated fetal outcome among women who had hyperemesis, with conflicting results (Depue et al. 1987; Kallen 1987; Godsey and Newman 1991; Hallak et al. 1996; Gross et al. 1989). Yet, a recent review states that severe vomiting does not have a negative effect on perinatal outcome (Eliakim et al. 2000). However, if the definition of the case series is based on a hospital diagnosis without taking severity into consideration, then the cases may include a number of women with a relatively mild disease that is not really distinct from the "normal" nausea and vomiting of pregnancy. Restricting cases to those where some objective biomarkers, such as severe ketosis or serum electrolyte disturbance, can be measured may then provide a "purer" case series, which might be one of the reasons for the inconsistent findings.

Timing of the disease could also be of relevance, as in two studies investigating the association between hyperemesis and female sex of the baby, it was noted that the association between hyperemesis and having a female baby was evident only for hospitalized cases of hyperemesis occurring in the first trimester of pregnancy (Askling et al. 1999; Basso et al. 2001). Hyperemesis occurs more frequently in women carrying multiple fetuses as well as in women (carrying singletons) who will later be diagnosed with pre-eclampsia (Zhang and Cai 1991). These observations suggest that there might be multiple causal paths leading to hyperemesis and that, in some women, a large placenta (with a high hormone production) may be the cause of the disease, while in others hyperemesis could be a sign of some other pathological process under way. In any event, whenever etiological heterogeneity exists in a process, and signs and symptoms are the same as that of another process (so that they are considered the same disorder), it is always very difficult to disentangle what is being studied. If only a small fraction of the case population represents a different etiological entity, even a moderately strong predictor of the “minority” disease entity will likely appear not to be associated with the mixed entity and will thus be repeatedly missed. If the fraction is large, then a predictor will be called a risk factor for all entities presenting with the same symptoms, and perhaps this uncertainty contributes to some of the failures of epidemiology to encourage changes in people’s habits and in policies. When these conditions occur in pregnancy, it may, in some cases, be possible to discriminate to some extent between different entities by paying particular attention to the timing in which the events occur.

45.9.3 Gestational Diabetes

As many other metabolic functions, the metabolism of carbohydrates is altered in normal pregnancies. The fasting blood glucose level decreases early in pregnancy until the 12th week and usually remains at this lower level until the end of pregnancy. Insulin, by contrast, remains stable during the first and second trimesters but increases during the third. Outside pregnancy, blood glucose returns rapidly to fasting levels after a meal, while, in pregnancy, both glucose and insulin levels reach higher peaks than they would after a similar meal in the non-pregnant state. This higher level is, furthermore, maintained for a longer time. Human placental lactogen (hPL) is the hormone that is believed to be responsible for these changes in metabolism (Haig 1993). In general, pregnancy is often a state of mild insulin resistance, and some women develop gestational diabetes. As an adaptation, insulin production is increased at the same time that the mother is becoming insulin resistant. Within his evolutionary theory about the genetic conflicts of pregnancy, Haig (1993) proposes an interesting hypothesis about why this change may occur. Glucose is an important nutrient for the growing fetus, but it is also important for the mother’s survival that not too many of her resources are depleted by the fetus, which could happen if fetal demands went unopposed: the decline in blood glucose early in pregnancy could thus represent a maternal attempt to limit fetal uptake. In addition, after every meal, there will be a competition between the mother and the fetus over the respective share. The longer it takes the mother to reduce her blood sugar, the higher the share taken by the fetus, hence the insulin resistance, according to Haig. At the beginning of pregnancy, the fetus has limited demands and limited “power,” which is why the mother succeeds in “hiding” her blood glucose. During the third trimester, however, the fetus has very high demands (as this is the time when most fetal weight is gained), and it is strong enough to take the upper hand in the competition with the mother. When seen in the light of how to optimize survival probabilities, this is an attractive hypothesis, although hard to test.

Barker (1995, 1998) has a different hypothesis, suggesting that insulin resistance after birth is a consequence of limited nutrient supply in fetal life. Clearly, however, glucose metabolism is crucial during pregnancy, and a mechanism has evolved that creates a delicate balance between the maternal and fetal needs. Diabetes during pregnancy is a complex problem that requires a very careful management to prevent damages to the fetus and to the mother. Preexisting diabetes is a risk factor for several congenital abnormalities, stillbirth, and macrosomia. Gestational diabetes is not associated with an increased risk of congenital abnormalities (likely because it occurs later in pregnancy), but the risk of stillbirth and macrosomia is present also for this condition (Schmidt et al. 2001; Johnstone et al. 1990). Macrosomia is the most frequent outcome in diabetic mothers, and this can complicate delivery to the point that a caesarean section is required. Women who have had gestational diabetes are at increased risk of having it again in a subsequent pregnancy and are also at a high risk of developing diabetes (Dornhorst and Beard 1993), especially type 2 diabetes (O'Sullivan and Mahan 1964; O'Sullivan 1991), later in life. Gestational diabetes, as well as other types of diabetes, is also a risk factor for pre-eclampsia (Schmidt et al. 2001), another potentially severe pregnancy complication. Obesity is a predisposing factor (with insulin resistance as the underlying condition), as is advanced maternal age or having a family history of diabetes.

The definition of gestational diabetes is problematic and the subject of controversy (Dornhorst and Beard 1993; Martin et al. 1995; Gabbe 1998; Bonomo et al. 1998; Schmidt et al. 2000). The discussion has to do with whether there should be universal screening for all pregnant women or only for those at higher risk. Risk factors for gestational diabetes are advanced maternal age, high body mass index, family history of diabetes, and certain ethnicities. There is no agreement over the "gold standard" test for women who are screened positive. The oral glucose tolerance test is the norm, but the load (the dose of administered glucose) varies geographically. There is also discussion about which cut-off limits should be used, as the prevalence of gestational diabetes would change depending upon the threshold, and the risk would be to either classify too many women who are not diabetic as such or to miss too many women who are diabetic and at risk of adverse pregnancy or health outcome. In a randomized trial among women with gestational diabetes of routine vs. nutritional advice, blood glucose monitoring, and insulin therapy as needed, Crowther et al. (2005) reported a significantly higher rate of perinatal complications (death, bone fracture, shoulder dystocia, and nerve palsy) among women in the routine care group vs. the intervention group.

The distinction between normal and pathological values is arbitrary, which hampers any diagnostic criterion. Often, epidemiologists do not deal with the actual values of glucose level but instead with the clinical diagnosis collected in several hospitals, without guarantee of uniformity in the criteria used for screening and diagnosis. The diagnosis will also depend on whether glucose is measured or not, and, since gestational diabetes is mostly asymptomatic, this is an added complication. Researchers planning studies requiring an accurate diagnosis of gestational diabetes should thus be aware of the medical attitude towards screening in pregnancy in the locations where they plan to collect their data and of the tools and cut-off levels in use as well as of the criteria that govern which women are screened and which are not. At best, studies are based upon follow-up of cohorts that are all subject to testing within the same protocol.

Geographical variations in the definition and incidence as well as changes in definition and screening attitudes over time are also to be taken into consideration when making comparisons between places and periods. Long- and short-term consequences for the baby have been identified, and in some countries, the focus of the diagnosis of gestational diabetes has now shifted from the likelihood of progression to later chronic diabetes in the mother to the outcome of pregnancy, a shift that also has consequences on the diagnostic criteria. Since obesity is associated with a highly increased risk of type 2 diabetes as well as gestational diabetes, a number of women would only be diagnosed during pregnancy, and the two types of diabetes would then be confused. However, since both types of diabetes increase the risk of pregnancy complications (such as pre-eclampsia) and adverse fetal outcome, this may not be a major problem, depending on the specific purpose of the study. When studying pre-eclampsia, for instance, there are situations where it would be advisable to exclude women with preexisting diabetes but not those with gestational diabetes (which shares with pre-eclampsia obesity as a risk factor and, possibly, other predictors), and this may prove difficult to do. If and how much of an impact this could have on the estimates will, once again, depend on the specific situation and will in many cases be hard to evaluate.

45.9.4 Pregnancy-Induced Hypertension and Pre-Eclampsia

45.9.4.1 Definition and Diagnosis

In the first trimester of pregnancy, blood pressure is usually reduced from normal values. In many women, however, blood pressure increases around mid-pregnancy to values above normal. A modest degree of hypertension is thus a rather common condition of pregnancy and has not been consistently associated with unfavorable outcomes.

Pre-eclampsia is, on the other hand, one of the most common and potentially severe complications of pregnancy. In pre-eclampsia, maternal blood pressure can increase dramatically, and the heart, brain, and kidneys may be severely damaged. If the mother survives, the affected organs usually return to normality shortly after delivery, but long-term morbidity can persist, and pre-eclampsia can be fatal. If seizures occur, the disease is called eclampsia (a very rare occurrence in countries with well-functioning health-care systems) and the risk for both the mother and the fetus is then much higher. While eclampsia is a dramatic event that is probably rarely misclassified, pre-eclampsia is, by definition, much more elusive. In most countries, it is currently defined as the concurrent presence of hypertension and proteinuria. Gestational hypertension is defined as either a persistent rise of 25 millimeter mercury in systolic blood pressure during pregnancy compared to pre-pregnant values (if the pre-pregnant values are not known, a systolic blood pressure of 140 mmHg or higher) or as a rise of 15 in diastolic blood pressure (DBP) (or as a DBP above 90). The definitions of gestational hypertension do, however, vary geographically. For the disorder to be called pre-eclampsia, hypertension must be accompanied by a specified level of proteinuria. The degree of severity depends on the values of blood pressure and the amount of protein loss, as well as on additional signs and symptoms, often including edema. Previously, pre-eclampsia was defined by the concomitant presence of two out of three symptoms (hypertension, proteinuria, and edema), but the definition has since changed to be restricted to cases where both hypertension and proteinuria are present at the same time, as edema was too unspecific. The problem with this definition is, however, that a mild state of hypertension is common in pregnancy and, often, the pre-pregnant values are not known. What is termed mild pre-eclampsia may thus, in some cases, be nothing more than a change in values of blood pressure and proteinuria within the norm. Sometimes, changes in these values may be severe enough to qualify for the diagnosis, but they will escape detection. Furthermore, some women become nervous when their blood pressure is taken in a clinical setting and may thus end up being classified as hypertensive due to a temporary and harmless rise in blood pressure ("white coat" hypertension). On the other hand, a number of cases may be missed by applying the definition, either because of ignorance of the baseline pre-pregnancy values or because women do not have their blood pressure measured at the moment of the increase, and, if there are no severe symptoms, the diagnosis will never be made. In severe cases, women become very sick and there is little doubt about the diagnosis, but these cases are the minority.

The reported incidence of pre-eclampsia appears to vary widely across places, from an estimated 2% to approximately 8%. This variation may reflect real variations in susceptibility and determinants across populations, but it almost certainly also depends on the sources of information for the diagnosis as well as on the access to prenatal care and the problems mentioned above.

The only known "cure" for pre-eclampsia is to end the pregnancy, as the placenta appears to be the organ that causes the disease, and pre-eclampsia is therefore the major cause of iatrogenic preterm delivery.

45.9.4.2 Pathophysiology of Pre-Eclampsia

In normal pregnancy, the maternal spiral arteries are modified and penetrate deeply into the decidua (first trophoblastic invasion) and, around the 16th to 18th week, into the myometrium (second trophoblastic invasion). The invasive trophoblast enlarges the vessels from within, and a fibrin substance that renders the vessels flaccid and unresponsive to maternal vasoconstriction replaces the vessels' internal lining. In pre-eclampsia often, but not always, the second trophoblastic invasion does not occur, or occurs only to a very modest degree (Salas 1999; Roberts and Lain 2002), resulting in placental perfusion being severely compromised because the arteries are narrow and with a high resistance, instead of being wide, low-resistance vessels, as they would be if the invasion had proceeded normally.

Haig (1993) expresses the view that hypertension in pregnancy is a fetal adaptive mechanism. Because of the structure of the modified spiral arteries, maternal ability to control the blood flow to the placenta is limited, and the placental site is characterized by low resistance to blood flow. Thus, for any given resistance of the placental unit, a compensatory rise in the maternal peripheral blood pressure will increase the blood flow to the placenta. According to Haig, this may be a sign of a feto-maternal conflict, where the growing fetus is able – by some unknown mechanism – to increase maternal blood pressure and thus increase the placental blood flow. Drug-induced reduction of mean arterial pressure may be associated with a reduction in fetal growth (von Dadelszen et al. 2000).

Pre-eclampsia is not, however, always accompanied by defective placentation and is, most likely, a common syndrome resulting from heterogeneous causes (Ness and Roberts 1996). It is believed that large placental mass (as seen in multiple pregnancies) and endothelial disease (as seen in diabetics) are mechanisms that can also produce placental hypoperfusion and start the cascade of events that leads to pre-eclampsia (Salas 1999).

45.9.4.3 Known Predictors of Pre-Eclampsia

The etiology of pre-eclampsia is mostly unknown, although understanding of the pathophysiology of pre-eclampsia has progressed dramatically in the last few decades. This disorder is one of the most tantalizing mysteries of reproductive epidemiology. The best-known predictors are multifetal pregnancies, nulliparity, obesity, and some maternal disease (such as kidney disease or diabetes), while smoking is protective for reasons unknown, although several hypotheses have been raised to explain this association (Condé-Agudelo et al. 1999). Africans and African Americans appear to be at a higher risk, possibly because susceptibility to pre-eclampsia is related to susceptibility to cardiovascular disease (Roberts and Lain 2002). One study suggested that women giving birth preterm with pre-eclampsia were at a higher risk of later death from cardiovascular disease, while women giving birth at term with pre-eclampsia had no increased risk compared with non-pre-eclamptic women (Irgens et al. 2001). A meta-analysis of studies of cardiovascular *sequelae* in women who have experience pre-eclampsia suggested an elevated risk, which further increased with the severity of pre-eclampsia (McDonald et al. 2008). Whether pre-eclampsia is an early manifestation of a predisposition to cardiovascular disease or, instead, it is a contributing cause is not clear.

Several trials have addressed the association between pre-eclampsia and dietary factors, mostly calcium, magnesium, antioxidants, and fish oil. Unfortunately, no clear answer has emerged from these studies, except for the finding that calcium appears to be protective among women with a very low baseline intake or for women with a very high risk of pre-eclampsia (Villar and Belizan 2000). In general, however, the attempts to prevent pre-eclampsia through dietary supplements or aspirin have been overall disappointing (Sibai 1998; Dekker and Sibai 2001). A Cochrane review in 2006 examined trials where at least 1g daily of calcium was administered to pregnant women, who had a generally low calcium intake, and indicated that calcium reduced the risk of pre-eclampsia as well as the risk of a composite outcome (death or serious morbidity) (Hofmeyr et al. 2006).

It is well accepted that a genetic component to pre-eclampsia exists, since children born of pre-eclamptic pregnancies are themselves at a higher risk of having children born of pregnancies with pre-eclampsia (Esplin et al. 2001; Skjaerven et al. 2005). Also, among males who have a child with a woman different from the one with whom they had previously had a child have almost twice the risk of having their subsequent partner also developing pre-eclampsia, compared to males whose previous partner had not had pre-eclampsia (Lie et al. 1998).

A large number of biomarkers and genetic factors have been explored as predisposing to pre-eclampsia (Broughton Pipkin 1999; Roberts and Cooper 2001). Genetic studies on pre-eclampsia have not consistently revealed a specific genotype associated with pre-eclampsia, although women with pre-eclampsia are more likely to have a heterozygous factor V Leiden mutation and other thrombophiliac mutations (Alfirevic et al. 2002). Not all researchers agree on the role of thrombophiliac mutations (Livingston et al. 2001), however.

Reduced levels of placental growth factors (PIGF) and elevated levels of markers predictive of pre-eclampsia, soluble fms-like tyrosine kinase 1 (sFlt-1), which binds to the placental growth factor (PlGF), predict the subsequent development of pre-eclampsia (Levine et al. 2004). Soluble endoglin, another antiangiogenic protein, acts together with sFlt-1 to induce a pre-eclampsia-like syndrome in pregnant rats, and a study indicated that the risk of pre-eclampsia was greater among women in the highest quartile for both biomarkers, but among those in the highest quartile for either biomarker alone (Levine et al. 2006). Soluble Flt-1 (sFlt-1) causes endothelial dysfunction by antagonizing vascular endothelial growth factor (VEGF) and placental growth factor (PlGF). VEGF maintains endothelial health in several tissues, including the kidney, and, in normal pregnancy, the placenta produces only modest concentrations of VEGF, PlGF, and soluble Flt-1 (Karumanchi et al. 2005). Makris et al. (2007) reported a syndrome analogous to human pre-eclampsia and an increase in circulating sFlt-1 in non-human primates after induction of utero-placental ischemia, suggesting that placental ischemia is sufficient to induce pre-eclampsia as well as elevated sFlt-1 (Karumanchi and Epstein 2007). Interestingly, sFlt-1 levels appear to be lower throughout pregnancy in smokers compared with non-smokers, and smokers also differed from non-smokers between the 10th and 20th week of gestation in levels of PIGF (higher) and soluble endoglins (lower), so the observed protective effect of smoking on pre-eclampsia may occur through some effect of nicotine on angiogenic proteins (Levine et al. 2006).

45.9.4.4 Methodological Challenges in Studies of Pre-Eclampsia

Although pre-eclampsia is probably the most studied among all pregnancy complications and keeps fascinating researchers from many areas of medicine, several difficulties face the investigators, mainly because of the difficulty of accurately identifying cases in sufficient numbers, without incurring selection problems. Research based upon nationwide hospital registries can provide population-based data that may, however, be of limited quality if the only available information is the code according to the International Classification of Diseases. The advantages of these studies are that women are most likely unselected and that the numbers will be large enough to allow studying even relatively rare predictors or outcomes. In some cases, these studies might be the only viable option. If, to improve the quality of the data, researchers restrict their case series to severe pre-eclampsia only, then the numbers will be dramatically reduced, but, possibly, fewer false positives will be included.

On the other hand, studies of the case-control type where medical charts are available would provide a much better case series if proper diagnostic procedures can be applied to document the disease, whereas problems may exist in recruiting a sufficient number of controls retaining a sufficient confidence that self-selection will not bias the study. If the women who accept to enter the study as controls do so according to the exposure under study, this will produce biased estimates to an extent that is often impossible to judge. Since pregnant women are invited to lead a healthy lifestyle for the sake of the baby if not their own, it is likely that some women whose pregnancy went well but whose habits were not "beyond reproach" would be relatively unwilling to take part in a study where such behaviors would be questioned. On the other hand, women who had a negative experience may be less reluctant to be under scrutiny, because they want to know what went wrong. However, problems in studying pre-eclampsia go well beyond the objective difficulties of appropriately defining cases or of finding unselected study populations.

Pre-eclampsia is a cause of preterm delivery, mostly iatrogenic. This fact complicates the interpretation of studies attempting to evaluate whether pre-eclampsia is associated with conditions that are more common in babies that are born preterm, such as cerebral palsy. Some studies have reported that babies of pre-eclamptic pregnancies were protected from cerebral palsy when the risk was examined by gestational week at birth (Gray et al. 1998; Murphy et al. 1995). Is this a protection conferred by pre-eclampsia, or is it an artefact due to the fact that the causes of preterm delivery in pre-eclamptic pregnancies differ from those of other preterm deliveries, where the causal factors may more frequently also be a cause of cerebral palsy? Because babies born preterm for causes other than pre-eclampsia have different pathologies that resulted in their early birth, disentangling the effects of preterm birth from its causes is a major challenge, as is examining the various causal paths leading to preterm birth that may very well be implicated in the diseases "associated" with preterm birth. Also, many cases of pre-eclampsia are delivered by emergency caesarean section, and the delivery complications due to caesarean section are generally different from those arising from vaginal deliveries, as several spontaneous preterm births will be. If complications that can arise from vaginal delivery were associated with cerebral palsy or other health problems (e.g., anoxia), then comparison of babies born of pre-eclamptic pregnancies with babies born of non-pre-eclamptic ones will be problematic.

Another problem has to do with studying pre-eclampsia in connection with other conditions or factors that are associated with preterm delivery. If a given factor causes preterm delivery, it may also appear to protect from pre-eclampsia simply because women with a shortened pregnancy have had less opportunity of developing it, since pre-eclampsia often occurs after the 36th week of gestation, but being pregnant (or just delivered) is a necessary condition for being diagnosed with it. If the date when pre-eclampsia was first diagnosed is known, data may be analyzed through Cox regression or survival methods to overcome this problem (for a general introduction to survival analysis, see chapter ►Survival Analysis of this handbook).

Beyond preterm birth, babies of pre-eclamptic pregnancies are at higher risk of fetal and neonatal death. However, the risk of stillbirth has declined over time, probably due to increased monitoring of pregnancies and earlier obstetric intervention. At least in Norway, however, the risk of neonatal death has increased only slightly among first-born babies of pre-eclamptic pregnancies, despite a marked increase in the rate of babies born before 32 weeks among pre-eclamptic mothers (Basso et al. 2006b). Babies of pre-eclamptic pregnancies are usually smaller on average than babies of normotensive pregnancies, at least if born by 37 weeks, while they are of comparable (or even larger) size thereafter (Xiong et al. 2002).

If an important confounder is systematically omitted when studying a given disease, this will lead to the potential establishment of a wrong conclusion (cf. chapter ▶Confounding and Interaction of this handbook). For example, a widespread hypothesis about the etiology of pre-eclampsia proposes that a maternal immune reaction to paternal antigens could be a cause of the failed trophoblastic invasion. This hypothesis was mainly based on the observation that pre-eclampsia is more frequent in first pregnancies. Among multiparous, women who had changed their partner from the previous pregnancy had an increased risk (Dekker et al. 1998; Dekker 1999; Trupin et al. 1996; Lie et al. 1998; Li and Wi 2000). Also, women with a long period of sexual cohabitation prior to a pregnancy and women using oral contraceptives appeared to be at a lower risk of pre-eclampsia than women with a short cohabitation period or those using barrier contraceptive methods (Robillard et al. 1994; Dekker et al. 1998). This suggested that a prolonged exposure to the partner's sperm may confer a protection that would reduce the risk of pre-eclampsia, thus leading to the notion that "primipaternity" was a risk factor for pre-eclampsia. For some researchers (including the authors of this chapter), however, this hypothesis has lost its plausibility since three studies (two based on the Norwegian Birth Registry (Skjaerven et al. 2002; Trogstad et al. 2001) and one based on a sample from the Danish Birth Registry (Basso et al. 2001)) independently reported that the increased risk of pre-eclampsia with change of partner disappeared if the interval between births was adjusted for. Women who change partner have, on average, a much longer interval between births: if any factor correlated with time has an impact on the risk of pre-eclampsia, then women waiting a longer time will have an increased risk, regardless of whether they change partner. This was found to be true in the abovementioned studies, even after maternal age was controlled for. This finding prompted a further study where time to pregnancy was investigated in association with pre-eclampsia, as a fraction of the women waiting a long time between pregnancies may be subfecund. Time to pregnancy, as previously mentioned, is a proxy for the couple's fecundity and thus a relatively crude marker, since it reflects a multitude of disorders. It is, however, interesting that an association between long time to pregnancy and pre-eclampsia could be observed despite these limitations (Basso et al. 2003), and this evidence may lead to further research for identifying a subgroup of infertile women with a specific disorder that relates to pre-eclampsia.

Pre-eclampsia is most likely the result of an interaction between the maternal and the fetal systems, but its diagnosis relies exclusively on symptoms that are observed in the mother. It is perhaps for this reason that, so far, pre-eclampsia has eluded most attempts to clarify its etiology, despite substantial progress in the understanding of its physiopathology in recent years.

45.10 Delivery Complications

Delivery complications may arise before or during delivery and present a risk for the mother and/or the baby. Some of the risks have to do with the fetus's presentation or its inability to pass through the birth canal. Unlike the other great apes, humans have very difficult deliveries, largely due to the tight fit of the infant's head in the mother's birth canal. This difficulty has to do with the rearrangements of the pelvis in humans to accommodate bipedalism as well as with the disproportionately large head of human infants (Trevathan). Until less than a century ago, obstructed labor was the major cause of fetal and maternal morbidity and mortality. Because of malnutrition, many women had underdeveloped pelvises and the baby's head would remain trapped in the birth canal. Nowadays, fetal or maternal death because of this is a very rare event in industrialized countries but still a major problem in developing countries, where most babies are delivered at home and hospitals are far away and may lack adequate resources. The three major causes of maternal mortality in developing countries are hemorrhage, sepsis, as well as hypertensive disorders, and the first two usually result from delivery complications.

Beyond fetal presentation (and position) and feto-pelvic disproportion, delivery complications also include weak contractions, prolonged labor (of any of the three stages), prolapse of the umbilical cord, perineal or vaginal tears, fetal asphyxia, retention of the placenta, and hemorrhage. Caesarean sections, which account for between 15% and 30% of all births, depending on countries, constitute perhaps the major difficulty when studying delivery complications. Caesarean sections can be planned or acute, and the latter could be started before delivery or during delivery. Emergency caesarean sections, themselves considered a "delivery complication," are triggered by complications arising in the mother or the fetus.

In the case of delivery complications even more than in other cases, researchers have to study what is left after physicians have acted. Therefore, only babies being born vaginally will be at risk of having specific accidents during their descent through the birth canal, accidents that may affect the supply of oxygen to the brain. This would not be a problem if the decision of delivering a woman by caesarean section were independent of any factors that may put the baby at higher risk of encountering such mishaps, but – usually – this is not the case. Since the relative size of the mother and the baby or signs of fetal distress may well trigger the decision of performing a caesarean section, it is likely that babies born vaginally and those born by caesarean section are not completely comparable before delivery, which will complicate any interpretation of findings associated with a given delivery complication. This will also complicate any study trying to evaluate the "effects" of any given intervention during delivery, as it will be difficult to separate the effects of the intervention from the causes that provoked it, which may also be the causes of the outcome of interest. Even restricting to planned caesareans may not be sufficient to solve the problem, as caesareans are planned for a reason, and a likely reason is that a complication is foreseen and a caesarean section may prevent it from occurring. Experience in previous pregnancies will also affect the mode of delivery. A woman who has previously delivered by caesarean section will, in many cases, have one also for her next delivery, especially if the two pregnancies are close in time. Since many events tend to repeat themselves in one woman's reproductive life, it will be hard to decide how to consider such women in a study, especially if the cause for the previous caesarean is not known.

A caesarean section is the preferred choice of delivery mode for an increasing number of women, and immediate risks appear to be few. Only little is, however, known about long-term effects, and recent results suggest that the risk of asthma may be increased (Kero et al. 2002), although recent findings have shown conflicting results.

Obstetric complications have been associated with schizophrenia (Geddes and Lawrie 1995; Verdoux et al. 1997), and hypoxia correlates with cerebral palsy (Blair and Stanley 1993). Anoxia or hypoxia will most likely cause cerebral damage, and it is reasonable to assume a causal link. But it may also be argued that a baby who had brain damage (which will later cause cerebral palsy) prior to birth will be more likely to have a complicated delivery and suffer from anoxia.

In general, if one wishes to study a delivery complication that may, in some cases, lead to a caesarean section (or to induction of delivery), it will be necessary to have information on why the caesarean section was performed. Practice of caesarean section, induction of delivery, instrumental birth, etc., change between geographical areas and in time, and dealing with these variations may prove a daunting and perhaps impossible task. Usually, many elements are used in the decision to treat and intervene, and it may just be impossible to identify all these elements and appropriately control for them in the analysis. Confounding by indication is one of the strongest arguments for evaluating treatments in randomized trials, which, however, will often be difficult to carry out in this context.

Since the practice of inducing birth (also by means other than caesarean section) is now widespread, with criteria for induction that often differ from one hospital to the other, it will always be difficult to study either induction itself or phenomena such as postterm delivery or macrosomia, even when good information about the causes of the induction are present.

If preterm babies are at a higher risk of incurring delivery complications, it may be this latter fact rather than the timing of birth that makes them at higher risk of several diseases. On the other hand, if some babies who are born preterm are born early because of some damage that will later cause the disease and makes them at higher risk of delivery complications, then delivery complications will spuriously appear to cause the disease.

45.11 Fetal Origins of Adult Diseases

Most reproductive epidemiology has been related to the time period from pregnancy planning to the early time period of a new life. Many diseases are now seen as trajectories that start at the time of conception, during pregnancy, or in early childhood. Obviously, studying exposures with an induction and latency time of causation spanning several decades raises severe problems of being able to control for intervening factors. Without longitudinal readings of the occurrence of possible confounding factors that may be affected by the exposure of interest, such studies may yield confounded results.

It is not unexpected (or even questionable) news that exposures during fetal life may have lasting effects. What is new is that prenatal exposures may cause diseases that surface to clinical detection long after birth, perhaps even as late as in the following generations. Fetal programming is the name that has been used to describe what could happen if a stimulus or insult at a critical period of organ development interferes with cell division and thus with the function of the affected organ. Permanent changes of organ function could, in principle, lead to many diseases (Olsen 2000), but best documented are the associations between origins of disproportional fetal growth and cardiovascular diseases, perhaps through insulin resistance (Barker 1994, 1995, 1998).

Although the brain is, to some extent, spared in periods of undernutrition, specific nutritional factors, stress, medicine, etc., may influence brain function. Studying determinants of brain function and brain pathology in fetal life is a high priority research area.

Hormonal factors during fetal life may not only affect the reproductive organs but could also be associated with other diseases, such as mental disorders or cancer of the breast, prostate, and testis. As early as 1990, Trichopoulos suggested that breast cancer might originate in utero. It was suggested that estrogen could play a role, and, since estrogen correlates with fetal growth, it is expected that rapid fetal growth could be associated with a higher risk of breast cancer five or more decades later. The hypothesis has, to some extent, been corroborated (McCormack et al. 2003), but we still lack cohort data of sufficient follow-up time with valid exposure data from the time of conception.

The role of infections during pregnancy is also an area of intense research activity, especially related to mental and neurological disorders (Xiao et al. 2009; Zhu 2009; Mortensen et al. 2007; Sørensen et al. 2009; Sun et al. 2008).

45.12 Sources of Data

As in other areas of epidemiology, data come from different sources: secondary routine data, or ad hoc data based upon self-reported information, information from clinical measures, or information extracted from biological samples – blood, urine, placenta tissue, etc. (Longnecker et al. 2001).

More secondary data are available in reproductive health than in most other epidemiological areas. Most pregnant women and most newborn children are carefully monitored, and data are stored in medical records or even in computerized birth registers that may include not only birth endpoint data but also exposure data such as smoking and medical treatment (Ericson et al. 1999).

The data that usually have to be collected for research are those that describe putative causes of reproductive failures. Many of these exposures have to be collected prospectively, since they are often forgotten or cannot be reconstructed in an unbiased way back in time, once the outcome of the pregnancy is known. This is often true for dietary factors, medical treatments, occupational exposures, etc. Data on infections, occupational exposures, lifestyle factors, dietary factors, etc., cannot be recalled for more than a few weeks or perhaps months, unless they tend to change little over time. Usually, the mother is, not unexpectedly, a better source for data on pregnancy and birth than the father (Coughlin et al. 1998).

For these and other reasons, it seems justified to set up large cohort studies, starting shortly after conception and with the aim of collecting exposure information during pregnancy. It is also necessary to establish cohorts that can be followed over long time periods, the best case scenario being from conception to death, including information on the offspring of the children born into the cohort. These cohorts need to be large to provide sufficient information for rare outcomes, such as congenital abnormalities. Large cohorts of this type were set up in the past, and the best known is probably the National Collaborative Perinatal Project from the USA, started in the late 1950s, where more than 50,000 pregnant women were enrolled. The cohort aimed at studying obstetrical complications and the risk of cerebral palsy and other neurological disorders, although the cohort has served many other research purposes since then.

The Danish National Birth Cohort (DNBC) enrolled 100,000 pregnant women from 1996 to 2002 (Olsen et al. 2001; Statens Serum Institut 2013, www.dnbc.dk) and included data from interviews, registers, and self-administered questionnaires, together with blood from the mother and child stored in a biobank (Statens Serum Institut 2013, www.bsmb.dk). In this cohort, there is an ongoing follow-up of children and parents based upon record linkage. More detailed data are collected when the child reaches 7 and 11 years of age. Similar studies have been done in Norway and China and are being conducted in other countries (Birthcohorts.net 2011).

45.13 Conclusions

In this chapter, we have tried to highlight some of the main features of reproductive epidemiology as we see them. In particular, the fact that reproduction, even in our medicalized world, is a direct result of selective processes, which are still active. In addition, most of the events that we study in this area are what is left after the selection of couples who succeeded in conceiving, and further selection of those conceptions will progress to clinical recognition and, potentially, medical intervention, which may lead to anticipated delivery, termination of pregnancy, or treatment of a disorder. For this reason, denominators are usually unknown. Furthermore, the processes that occur during a pregnancy that ends in a birth are usually also hidden, and therefore we do not really know what has happened to the fetus during the most delicate phases of development.

Any event in reproduction generally concerns two individuals rather than one (three when counting the father). In many instances, pregnancies are voluntary events, and many women have more than one pregnancy, although the decision to have further pregnancies often depends to some extent on the outcome of the previous ones and almost all adverse outcomes have a relatively high recurrence risk. Time is also of crucial importance when dealing with reproductive epidemiology, but its dimension is generally different from the time involved in the development of, say, cancer after exposure to a given mutagenic substance. The types of bias that can occur in this area are, in many cases, peculiar to this discipline, and they have to be taken into consideration.

Genetic and functional genetic studies are playing an increasing role in this field of study. We know now that gene expression is modified by epigenetic changes that provide room for "programming" health later in life (Foley et al. 2009). Although we may not be able to answer the big questions, such as how the entire process of organ development is coordinated, finding answers to less broad questions is of interest. We need to know much more about genetic and gene-environment interaction, as well as epigenetics, not only in the development of congenital malformations and childhood cancers but also for long-term organ programming (for an introduction to genetic epidemiology, see chapter ▶Statistical Methods in Genetic Epidemiology of this handbook). Using information on genetic factors in, e.g., metabolism of environmental exposures, such as alcohol, may even be of help in examining confounding. How much of the association between, e.g., alcohol and reproductive failures is due to confounding cannot be examined in a randomized trial, but the genetic factors that modify alcohol metabolism may follow "Mendelian randomization" and thus provide a design for comparison that bypasses some of the problems associated with the intercorrelation between lifestyle factors (Smith and Ebrahim 2003, 2005).

The peculiarities of reproductive epidemiology offer a number of opportunities to researchers willing to exploit them. We have tried to introduce readers to a number of the features that make this area of epidemiology exciting, vibrating, and fairly unique.

References

AbouZahr C (1998) Maternal mortality overview. In: Murray CJL, Lopez AD (eds) Health dimensions of sex and reproduction. Harvard University Press, Boston, p 144

Adams MM, Berg CJ, Rhodes PH, McCarthy BJ (1991) Another look at the black-white gap in gestation-specific perinatal mortality. Int J Epidemiol 20(4):950–957

Adams M, Andersen AM, Andersen PK, Haig D, Henriksen TB, Hertz-Picciotto I, Lie RT, Olsen J, Skjaerven R, Wilcox A (2003) Sostrup statement on low birth weight (LBW). Int J Epidemiol 32(5):884–885

Alderete E, Eskenazi B, Sholtz R (1995) Effect of cigarette smoking and coffee drinking on time to conception. Epidemiology 6(4):403–408

Alfirevic Z, Roberts D, Martlew V (2002) How strong is the association between maternal thrombophilia and adverse pregnancy outcome? A systematic review. Eur J Obstet Gynecol Reprod Biol 101(1):6–14

Ananth CV, Demissie K, Smulian JC, Vintzileos AM (2001) Relationship among placenta previa, fetal growth restriction, and preterm delivery: a population-based study. Obstet Gynecol 98(2):299–306

Andersen AM, Vastrup P, Wohlfahrt J, Andersen PK, Olsen J, Melbye M (2002) Fever in pregnancy and risk of fetal death: a cohort study. Lancet 360(9345):1552–1556

Askling J, Erlandsson G, Kaijser M, Akre O, Ekbom A (1999) Sickness in pregnancy and sex of child. Lancet 354(9195):2053

Baird DD, Wilcox AJ (1985) Cigarette smoking associated with delayed conception. JAMA 253(20):2979–2983

Baird DD, Wilcox AJ, Weinberg CR (1986) Use of time to pregnancy to study environmental exposures. Am J Epidemiol 124(3):470–480

Baird DD, Ragan BN, Wilcox AJ, Weinberg CR (1993) The relationship between reduced fecundability and subsequent foetal loss. In: Gray R, Leridon H, Spira A (eds) Biomedical and demographic determinants of reproduction. Clarendon, Oxford, pp 329–341

Baird DD, Weinberg CR, Schwingl P, Wilcox AJ (1994) Selection bias associated with contraceptive practice in time-to-pregnancy studies. Ann N Y Acad Sci 709:156–164

Barker DJ (1994) Mothers, babies and disease in later life. BMJ Publishing Group, London

Barker DJ (1995) Fetal origins of coronary heart disease. BMJ 311(6998):171–174

Barker DJ (1998) Mothers, babies and health later in life, 2nd edn. Churchill Livingstone, Edinburgh

Basso O (2007) Options and limitations in studies of successive pregnancy outcomes: an overview. Paediatr Perinat Epidemiol 21(Suppl 1):8–12

Basso O, Baird DD (2003) Infertility and preterm delivery, birthweight, and Caesarean section: a study within the Danish National Birth Cohort. Hum Reprod 18(11):2478–2484

Basso O, Wilcox AJ (2009) Intersecting birth weight-specific mortality curves: solving the riddle. Am J Epidemiol 169(7):787–797

Basso O, Wilcox A (2010) Mortality risk among preterm babies: immaturity versus underlying pathology. Epidemiology 21(4):521–527

Basso O, Wilcox AJ (2011) Might rare factors account for most of the mortality of preterm babies? Epidemiology 22(3):320–327

Basso O, Olsen J, Knudsen LB, Christensen K (1998) Low birth weight and preterm birth after short interpregnancy intervals. Am J Obstet Gynecol 178(2):259–263

Basso O, Fonager K, Olsen J (2000a) Are pregnancies shorter in leap years? Epidemiology 11(6):736–737

Basso O, Juul S, Olsen J (2000b) Time to pregnancy as a correlate of fecundity: differential persistence in trying to become pregnant as a source of bias. Int J Epidemiol 29(5):856–861

Basso O, Christensen K, Olsen J (2001) Higher risk of pre-eclampsia after change of partner. An effect of longer interpregnancy intervals? Epidemiology 12(6):624–629

Basso O, Weinberg CR, Baird DD, Wilcox AJ, Olsen J (2003) Subfecundity as a correlate of preeclampsia: a study within the Danish National Birth Cohort. Am J Epidemiol 157(3):195–202

Basso O, Christensen K, Olsen J (2004a) Fecundity and twinning. A study within the Danish National Birth Cohort. Hum Reprod 19(10):2222–2226

Basso O, Nohr EA, Christensen K, Olsen J (2004b) Risk of twinning as a function of maternal height and body mass index. JAMA 291(13):1564–1566

Basso O, Frydenberg M, Olsen SF, Olsen J (2005) Two definitions of “small size at birth” as predictors of motor development at six months. Epidemiology 16(5):657–663

Basso O, Wilcox AJ, Weinberg CR (2006a) Birth weight and mortality: causality or confounding? Am J Epidemiol 164(4):303–311

Basso O, Rasmussen S, Weinberg CR, Wilcox AJ, Irgens LM, Skjaerven R (2006b) Trends in fetal and infant survival following preeclampsia. JAMA 296(11):1357–1362. Erratum in: JAMA 296(24):2926

Berkowitz GS, Papiernik E (1993) Epidemiology of preterm birth. Epidemiol Rev 15(2):414–443

Bianchi DW (2000) Fetomaternal cell trafficking: a new cause of disease? Am J Med Genet 91(1):22–28

Bianchi DW, Zickwolf GK, Weil GJ, Sylvester S, DeMaria MA (1996) Male fetal progenitor cells persist in maternal blood for as long as 27 years postpartum. Proc Natl Acad Sci USA 93(2):705–708

Birthcohorts.net (2011). http://www.birthcohorts.net/. Accessed 1 May 2013

Blair E, Stanley F (1993) When can cerebral palsy be prevented? The generation of causal hypotheses by multivariate analysis of a case-control study. Paediatr Perinat Epidemiol 7(3):272–301

Bolumar F, Olsen J, Boldsen J (1996) Smoking reduces fecundity: a European multicenter study on infertility and subfecundity. The European Study Group on Infertility and Subfecundity. Am J Epidemiol 143(6):578–587

Bonde JP, Ernst E, Jensen TK, Hjollund NHI, Kolstad H, Henriksen TB, Scheike T, Givercman A, Olsen J, Skakkebaek NE (1998a) Relation between semen quality and fertility, a population-based study of 430 first-pregnancy planners. Lancet 352(9135):1172–1177

Bonde JP, Hjollund NHI, Jensen TK, Ernst E, Kolstad H, Henriksen TB, Givercman A, Skakkebaek NE, Andersen AM, Olsen J (1998b) A follow-up study of environmental and biologic determinants of fertility among 430 Danish first-pregnancy planners: design and methods. Reprod Toxicol 12(1):19–27

Bonde JP, Joffe M, Sallmén M, Kristensen P, Olsen J, Roeleveld N, Wilcox A (2006) Validity issues relating to time-to-pregnancy studies of infertility. Epidemiology 17(4):347–349

Bonomo M, Gandini ML, Mastropasqua A, Begher C, Valentini U, Faden D, Morabito A (1998) Which cutoff level should be used in screening for glucose intolerance in pregnancy? Definition of Screening Methods for Gestational Diabetes Study Group of the Lombardy Section of the Italian Society of Diabetology. Am J Obstet Gynecol 179(1):179–185

Broughton Pipkin F (1999) What is the place of genetics in the pathogenesis of pre-eclampsia? Biol Neonate 76(6):325–330

Carlsen E, Giwercman A, Keiding N, Skakkebaek NE (1992) Evidence for decreasing quality of semen during past 50 years. BMJ 305(6854):609–613

Chalmers I (1979) The search for indices. Lancet 2(8151):1063–1065

Christensen K, Basso O, Kyvik KO, Juul S, Boldsen J, Vaupel JW, Olsen J (1998) Fecundability of female twins. Epidemiology 9(2):189–192

Cnattingius S, Signorello LB, Anneren G, Clausson B, Ekbom A, Ljunger E, Blot WJ, McLaughlin JK, Petersson G, Rane A, Granath F (2000) Caffeine intake and the risk of first-trimester spontaneous abortion. N Engl J Med 343(25):1839–1845

Cohen JM, Hutcheon JA, Kramer MS, Joseph KS, Abenhaim H, Platt RW (2010) Influence of ultrasound-to-delivery interval and maternal-fetal characteristics on validity of estimated fetal weight. Ultrasound Obstet Gynecol 35(4):434–441

Comstock GW, Lundin FE Jr (1967) Parental smoking and perinatal mortality. Am J Obstet Gynecol 98(5):708–718

Condé-Agudelo A, Althabe F, Belizan JM, Kafury-Goeta AC (1999) Cigarette smoking during pregnancy and risk of preeclampsia: a systematic review. Am J Obstet Gynecol 181(4): 1026–1035

Cooney MA, Buck Louis GM, Sundaram R, McGuinness BM, Lynch CD (2009) Validity of self-reported time to pregnancy. Epidemiology 20(1):56–59

Coughlin MT, LaPorte RE, O'Leary LA, Lee PA (1998) How accurate is male recall of reproductive information? Am J Epidemiol 148(8):806–809

Cramer DW, Goldman MB (guest eds) (1994) Study designs and statistics for infertility research. Infertility and reproductive medicine. Clinics of North America, vol 5(2). W. B. Saunders, Philadelphia

Crowther CA, Hiller JE, Moss JR, McPhee AJ, Jeffries WS, Robinson JS (2005) Effect of treatment of gestational diabetes mellitus on pregnancy outcomes. Australian Carbohydrate Intolerance Study in Pregnant Women (ACHOIS) Trial Group. N Engl J Med 352(24):2477–2486

Dashe JS, McIntire DD, Ramus RM, Santos-Ramos R, Twickler DM (2002) Persistence of placenta previa according to gestational age at ultrasound detection. Obstet Gynecol 99(5 Pt 1):692–697

Dekker GA (1999) Risk factors for pre-eclampsia. Clin Obstet Gynecol 42(3):422–435

Dekker G, Sibai B (2001) Primary, secondary, and tertiary prevention of pre-eclampsia. Lancet 357(9251):209–215

Dekker GA, Robillard PY, Hulsey TC (1998) Immune maladaptation in the etiology of preeclampsia: a review of corroborative epidemiologic studies. Obstet Gynecol Surv 53(6):377–382

Depue RH, Bernstein L, Ross RK, Judd HL, Henderson BE (1987) Hyperemesis gravidarum in relation to estradiol levels, pregnancy outcome, and other maternal factors: a seroepidemiologic study. Am J Obstet Gynecol 156(5):1137–1141

DeRoo LA, Wilcox AJ, Drevon CA, Lie RT (2008) First-trimester maternal alcohol consumption and the risk of infant oral clefts in Norway: a population-based case-control study. Am J Epidemiol 168(6):638–646

Diggle GE (2001) Thalidomide: 40 years on. Int J Clin Pract 55(9):627–631

Dornhorst A, Beard RW (1993) Gestational diabetes: a challenge for the future. Diabet Med 10(10):897–905

Draper ES, Kurinczuk JJ, Abrams KR, Clarke M (1999) Assessment of separate contributions to perinatal mortality of infertility history and treatment: a case-control analysis. Lancet 353(9166):1746–1749

Eliakim R, Abulafia O, Sherer DM (2000) Hyperemesis gravidarum: a current review. Am J Perinatol 17(4):207–218

Ericson A, Kallen B, Wiholm B (1999) Delivery outcome after the use of antidepressants in early pregnancy. Eur J Clin Pharmacol 55(7):503–508

Esplin MS, Fausett MB, Fraser A (2001) Paternal and maternal components of the predisposition to preeclampsia. N Engl J Med 344(12):867–872

Fei C, McLaughlin JK, Lipworth L, Olsen J (2009) Maternal levels of perfluorinated chemicals and subfecundity. Hum Reprod 24(5):1200–1205

Ferrari RM, Cooney MA, Vexler A, Liu A, Buck Louis GM (2007) Time to pregnancy and multiple births. Hum Reprod 22(2):407–413

Flint C, Larsen H, Nielsen GL, Olsen J, Sorensen HT (2002) Pregnancy outcome after suicide attempt by drug use: a Danish population-based study. Acta Obstet Gynecol Scand 81(6):516–522

Foley DL, Craig JM, Morley R, Olsson CA, Dwyer T, Smith K, Saffery R (2009) Prospects for epigenetic epidemiology. Am J Epidemiol 169(4):389–400

Fortes Filho JB, Costa MC, Eckert GU, Santos PG, Silveira RC, Procianoy RS (2011) Maternal preeclampsia protects preterm infants against severe retinopathy of prematurity. J Pediatr 158(3):372–376

Fugazzola L, Cirello V, Beck-Peccoz P (2010) Fetal cell microchimerism in human cancers. Cancer Lett 287(2):136–141

Gabbe SG (1998) The gestational diabetes mellitus conferences. Three are history: focus on the fourth. Diabetes Care 21(Suppl 2):B1–B2

Geddes JR, Lawrie SM (1995) Obstetric complications and schizophrenia: a meta-analysis. Br J Psychiatry 167(6):786–793

Gjessing HK, Skjaerven R, Wilcox AJ (1999) Errors in gestational age: evidence of bleeding early in pregnancy. Am J Public Health 89(2):213–218

Gladen BC (1986) On the role of “habitual aborters” in the analysis of spontaneous abortion. Stat Med 5(6):557–564

Gluckman P, Hanson M (2005) The fetal matrix: evolution, development and disease. Cambridge University Press, Cambridge

Godsey RK, Newman RB (1991) Hyperemesis gravidarum. A comparison of single and multiple admissions. J Reprod Med 36(4):287–290

Gourbin G, Masuy-Stroobant G (1995) Registration of vital data: are live births and stillbirths comparable all over Europe? Bull World Health Organ 73(4):449–460

Grandjean P, Weihe P, White RF, Debes F, Araki S, Yokoyama K, Murata K, Sorensen N, Dahl R, Jorgensen PJ (1997) Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury. Neurotoxicol Teratol 19(6):417–428

Gray PH, O'Callaghan MJ, Mohay HA, Burns YR, King JF (1998) Maternal hypertension and neurodevelopmental outcome in very preterm infants. Arch Dis Child Fetal Neonatal Ed 79(2):F88–F93

Greenwood C, Yudkin P, Sellers S, Impey L, Doyle P (2005) Why is there a modifying effect of gestational age on risk factors for cerebral palsy? Arch Dis Child Fetal Neonatal Ed 90:F141–F146

Gross S, Librach C, Cecutti A (1989) Maternal weight loss associated with hyperemesis gravidarum: a predictor of fetal outcome. Am J Obstet Gynecol 160(4):906–909

Haig D (1993) Genetic conflicts in human pregnancy. Q Rev Biol 68(4):495–532

Haig D (2003) Meditations on birth weight: is it better to reduce the variance or increase the mean? Epidemiology 14(4):490–492

Hallak M, Tsalamandris K, Dombrowski MP, Isada NB, Pryde PG, Evans MI (1996) Hyperemesis gravidarum. Effects on fetal outcome. J Reprod Med 41(11):871–874

Hansen M, Kirinczuk JJ, Bower C, Webb S (2002) The risk of major birth defects after intracytoplasmic sperm injection and in vitro fertilization. N Engl J Med 346(10):725–730

Hediger ML, Scholl TO, Schall JI, Miller LW, Fischer RL (1995) Fetal growth and the etiology of preterm delivery. Obstet Gynecol 85:175–182

Henriksen TB, Wilcox A (1994) Ultrasound dating subject to bias. BMJ 308(6922):201

Henriksen TB, Wilcox AJ, Hedegaard M, Secher NJ (1995) Bias in studies of preterm and postterm delivery due to ultrasound assessment of gestational age. Epidemiology 6(5):533–537

Henriksen TB, Baird DD, Olsen J, Hedegaard M, Secher NJ, Wilcox AJ (1997) Time to pregnancy and preterm delivery. Obstet Gynecol 89(4):594–599

Hernandez-Diaz S, Schisterman EF, Hernan MA (2006) The birth weight "paradox" uncovered? Am J Epidemiol 164(11):1115–1120

Hjollund NH, Jensen TK, Bonde JP, Henriksen TB, Andersson AM, Kolstad HA, Ernst E, Giwercman A, Skakkebaek NE, Olsen J (2000) Spontaneous abortion and physical strain around implantation, a follow-up study of first-pregnancy planners. Epidemiology 11(1): 18–23

Hoffman EB, Sen PK, Weinberg CR (2001) Within-cluster resampling. Biometrika 88(4): 1121–1134

Hofmeyr GJ, Atallah AN, Duley L (2006) Calcium supplementation during pregnancy for preventing hypertensive disorders and related problems. Cochrane Database Syst Rev 19(3): CD001059

Howards PP, Schisterman EF, Heagerty PJ (2007a) Potential confounding by exposure history and prior outcomes: an example from perinatal epidemiology. Epidemiology 18(5):544–551

Howards PP, Hertz-Picciotto I, Poole C (2007b) Conditions for bias from differential left truncation. Am J Epidemiol 165(4):444–452

Howards PP, Schisterman EF, Poole C, Kaufman JS, Weinberg CR (2012) "Toward a clearer definition of confounding" revisited with directed acyclic graphs. Am J Epidemiol 176(6): 506–511

Hutcheon JA, Platt RW (2008) The missing data problem in birth weight percentiles and thresholds for "small-for-gestational-age". Am J Epidemiol 167:786–792

Irgens HU, Reisaeter L, Irgens LM, Lie RT (2001) Long term mortality of mothers and fathers after pre-eclampsia: population based cohort study. BMJ 323(7323):1213–1217

Jacobsen R, Bostofte E, Engholm G, Hansen J, Olsen JH, Skakkebaek NE, Moller H (2000) Risk of testicular cancer in men with abnormal semen characteristics: cohort study. BMJ 321(7264):789–792

James WH (1982) Second survey of secular trends in twinning rates. J Biosoc Sci 14(4):481–497

James WH (1986) Recent secular trends in dizygotic twinning rates in Europe. J Biosoc Sci 18(4):497–504

Jensen TK, Henriksen TB, Hjollund NHI, Scheike T, Kolstad H, Giwercman A, Ernst E, Bonde JP, Skakkebaek NE, Olsen J (1998) Caffeine intake and fecundability. A follow-up study among 430 Danish couples planning their first pregnancy. Reprod Toxicol 12(3):289–295

Joffe M, Villard L, Li Z, Plowman R, Vessey M (1993) Long-term recall of time-to-pregnancy. Fertil Steril 60(1):99–104

Joffe M, Villard L, Li Z, Plowman R, Vessey M (1995) A time to pregnancy questionnaire designed for long term recall: validity in Oxford, England. J Epidemiol Community Health 49(3): 314–319

Johnstone FD, Nasrat AA, Prescott RJ (1990) The effect of established and gestational diabetes on pregnancy outcome. Br J Obstet Gynaecol 97(11):1009–1015

Joseph KS (2004) Incidence-based measures of birth, growth restriction, and death can free perinatal epidemiology from erroneous concepts of risk. J Clin Epidemiol 57(9):889–897

Joseph KS, Huang L, Liu S, Ananth CV, Allen AC, Sauve R, Kramer MS (2007) Fetal and Infant Health Study Group of the Canadian Perinatal Surveillance System. Reconciling the high rates of preterm and postterm birth in the United States. Obstet Gynecol 109(4):813–822

Juul S, Karmaus W, Olsen J (1999) Regional differences in waiting time to pregnancy: pregnancy-based surveys form Denmark, France, Germany, Italy and Sweden. The European Infertility and Subfecundity Study Group. Hum Reprod 14(5):1250–1254

Kacem Y, Cannie MM, Kadji C, Dobrescu O, Lo Zito L, Ziane S, Strizek B, Evrard AS, Gubana F, Gucciardo L, Staelens R, Jani JC (2013) Fetal weight estimation: comparison of two-dimensional US and MR imaging assessments. Radiology 267:902–910

Kallen B (1987) Hyperemesis during pregnancy and delivery outcome: a registry study. Eur J Obstet Gynecol Reprod Biol 26(4):291–302

Karumanchi SA, Epstein FH (2007) Placental ischemia and soluble fms-like tyrosine kinase 1: cause or consequence of preeclampsia? Kidney Int 71(10):959–961

Karumanchi SA, Maynard SE, Stillman IE, Epstein FH, Sukhatme VP (2005) Preeclampsia: a renal perspective. Kidney Int 67(6):2101–2113

Kero J, Gissler M, Gronlund MM, Kero P, Koskinen P, Hemminki E, Isolauri E (2002) Mode of delivery and asthma – is there a connection? Pediatr Res 52(1):6–11

Kiely JL, Paneth N, Susser M (1985) Fetal death during labor: an epidemiologic indicator of level of obstetric care. Am J Obst Gynecol 153(7):721–727

Kjaer D, Horvath-Puhó E, Christensen J, Vestergaard M, Czeizel AE, Sørensen HT, Olsen J (2007) Use of phenytoin, phenobarbital, or diazepam during pregnancy and risk of congenital abnormalities: a case-time-control study. Pharmacoepidemiol Drug Saf 16(2):181–188

Kjaer D, Horvath-Puhó E, Christensen J, Vestergaard M, Czeizel AE, Sørensen HT, Olsen J (2008) Antiepileptic drug use, folic acid supplementation, and congenital abnormalities: a population-based case-control study. BJOG 115(1):98–103

Klebanoff MA, Shiono PH (1995) Top down, bottom up and inside out: reflections on preterm birth. Paediatr Perinat Epidemiol 9:125–129

Klebanoff MA, Schoendorf KC (2004) Invited commentary: what's so bad about curves crossing anyway? Am J Epidemiol 160:211–212; discussion 215–216

Kline J, Stein Z, Susser M (1989) Conception to birth – epidemiology of prenatal development. Monographs in Epidemiology and Biostatistics, vol 14. Oxford University Press, New York

Knox EG, Armstrong EH, Lancashire R (1984) The quality of notification of congenital malformations. J Epidemiol Community Health 38(4):296–305

Kondo S, Schutte BC, Richardson RJ, Bjork BC, Knight AS, Watanabe Y, Howard E, de Lima RL, Daack-Hirsch S, Sander A, McDonald-McGinn DM, Zackai EH, Lammer EJ, Aylsworth AS, Ardinger HH, Lidral AC, Pober BR, Moreno L, Arcos-Burgos M, Valencia C, Houdayer C, Bahuau M, Moretti-Ferreira D, Richieri-Costa A, Dixon MJ, Murray JC (2002) Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. Nat Genet 32(2):285–289

Kramer MS, Liu S, Luo Z, Yuan H, Platt RW, Joseph KS (2002) Analysis of perinatal mortality and its components: time for a change? Am J Epidemiol 156(6):493–497

Levine RJ, Symons MJ, Balogh SA, Milby TH, Whorton MD (1981) A method for monitoring the fertility of workers. 2. Validation of the method among workers exposed to dibromochloropropane. J Occup Med 23(3):183–188

Levine RJ, Maynard SE, Qian C, Lim KH, England LJ, Yu KF, Schisterman EF, Thadhani R, Sachs BP, Epstein FH, Sibai BM, Sukhatme VP, Karumanchi SA (2004) Circulating angiogenic factors and the risk of preeclampsia. N Engl J Med 350(7):672–683

Levine RJ, Lam C, Qian C, Yu KF, Maynard SE, Sachs BP, Sibai BM, Epstein FH, Romero R, Thadhani R, Karumanchi SA, CPEP Study Group (2006) Soluble endoglin and other circulating antiangiogenic factors in preeclampsia. N Engl J Med 355(10):992–1005. Erratum in: N Engl J Med 355(17):1840

Li DK, Wi S (2000) Changing paternity and the risk of preeclampsia/eclampsia in the subsequent pregnancy. Am J Epidemiol 151(1):57–62

Lie RT, Rasmussen S, Brunborg H, Gjessing HK, Lie-Nielsen E, Irgens LM (1998) Fetal and maternal contributions to risk of pre-eclampsia: population based study. BMJ 316(7141): 1343–1347

Lie RT, Lyngstadaas A, Ørstavik KH, Bakketeig LS, Jacobsen G, Tanbo T (2005) Birth defects in children conceived by ICSI compared with children conceived by other IVF-methods: a meta-analysis. Int J Epidemiol 34(3):696–701

Livingston JC, Barton JR, Park V, Haddad B, Phillips O, Sibai BM (2001) Maternal and fetal inherited thrombophilias are not related to the development of severe preeclampsia. Am J Obstet Gynecol 185(1):153–157

Longnecker MP, Klebanoff MA, Zhou H, Brock JW (2001) Association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for-gestational-age babies at birth. Lancet 358(9276):110–114

Ludwig M, Diedrich K (2002) Follow-up of children born after assisted reproductive technologies. Reprod Biomed Online 5(3):317–322

Lummaa V, Pettay JE, Russell AF (2007) Male twins reduce fitness of female co-twins in humans. Proc Natl Acad Sci USA 104(26):10915–10920

Macklon NS, Geraedts JP, Fauser BC (2002) Conception to ongoing pregnancy: the ‘black box’ of early pregnancy loss. Hum Reprod Update 8(4):333–343

MacLehose RF, Kaufman JS (2012) Commentary: the wizard of odds. Epidemiology 23(1):10–12; discussion 13–14

Maclure M (1991) The case-crossover design: a method for studying transient effects on the risk of acute events. Am J Epidemiol 133(2):144–153

Macmahon B, Alpert M, Salber EJ (1965) Infant weight and parental smoking habits. Am J Epidemiol 82(3):247–261

Makris A, Thornton C, Thompson J, Thomson S, Martin R, Ogle R, Waugh R, McKenzie P, Kirwan P, Hennessy A (2007) Uteroplacental ischemia results in proteinuric hypertension and elevated sFLT-1. Kidney Int 71(10):977–984

Mann JR, McDermott S, Griffith MI, Hardin J, Gregg A (2011) Uncovering the complex relationship between pre-eclampsia, preterm birth and cerebral palsy. Paediatr Perinat Epidemiol 25(2):100–110

Martin FI, Ratnaike S, Wootton A, Condos P, Suter PE (1995) The 75 g oral glucose tolerance in pregnancy. Diabetes Res Clin Pract 27(2):147–151

McClure I (2003) The essential difference: men, women and the extreme male brain. BMJ 327(7405):57

McCormack VA, dos Santos Silva I, De Stavola BL, Mohsen R, Leon DA, Lithell HO (2003) Fetal growth and subsequent risk of breast cancer: results from long term follow up of Swedish cohort. BMJ 326(7383):248

McDonald SD, Malinowski A, Zhou Q, Yusuf S, Devereaux PJ (2008) Cardiovascular sequelae of preeclampsia/eclampsia: a systematic review and meta-analyses. Am Heart J 156(5):918–930

McElrath TF, Hecht JL, Dammann O, Boggess K, Onderdonk A, Markenson G, Harper M, Delpapa E, Allred EN, Leviton A, ELGAN Study Investigators (2008) Pregnancy disorders that lead to delivery before the 28th week of gestation: an epidemiologic approach to classification. Am J Epidemiol 168:980–989

Mednick SA, Huttunen MO, Machon RA (1994) Prenatal influenza infections and adult schizophrenia. Schizophr Bull 20(2):263–267

Meyer MB, Comstock GW (1972) Maternal cigarette smoking and perinatal mortality. Am J Epidemiol 96(1):1–10

Mitchell AA, Cottler LB, Shapiro S (1986) Effect of questionnaire design on recall of drug exposure in pregnancy. Am J Epidemiol 123(4):670–676

Mortensen PH, Nørgaard-Pedersen B, Waltoft BL, Sørensen TL, Hougaard D, Yolken RH (2007) Early infections of Toxoplasma gondii and the later development of schizophrenia. Schizophr Bull 33(3):741–744

Murphy DJ, Sellers S, MacKenzie IZ, Yudkin PL, Johnson AM (1995) Case-control study of antenatal and intrapartum risk factors for cerebral palsy in very preterm singleton babies. Lancet 346(8988):1449–1454

Ness RB, Grainger DA (2008) Male reproductive proteins and reproductive outcomes. Am J Obstet Gynecol 198(6):620.e1–e4

Ness RB, Roberts JM (1996) Heterogeneous causes constituting the single syndrome of preeclampsia: a hypothesis and its implications. Am J Obstet Gynecol 175(5):1365–1370

Nohr EH, Vaeth M, Baker JL, Sørensen TIA, Olsen J, Rasmussen KM (2008) Combined associations of prepregnancy body mass index and gestational weight gain with the outcome of pregnancy. AM J Clin Nutr 87(6):1750–1759

Nybo Andersen A-M, Wohlfahrt J, Christens P, Olsen J, Melbye M (2000) Maternal age and fetal loss: population based register linkage study. BMJ 320(7251):1708–1712

Olsen J (1984) Calculating risk ratios for spontaneous abortions: the problem of induced abortions. Int J Epidemiol 13(3):347–350

Olsen J (1994) Options in making use of pregnancy history in planning and analysing studies of reproductive failure. J Epidemiol Community Health 48(2):171–174

Olsen J (2000) Prenatal exposures and long-term health effects. Epidemiol Rev 22(1):76–81

Olsen J (2008) Confounding by exposure history and prior outcome. Epidemiology 19(4):635–636

Olsen J, Andersen PK (1998) Accounting for pregnancy dependence in epidemiologic studies of pregnancy outcomes. Epidemiology 9(3):363–364

Olsen J, Andersen PK (1999) We should monitor human fecundity, but how? A suggestion for a new method that may also be used to identify determinants of low fecundity. Epidemiology 10(4):419–421

Olsen J, Rachootin P (2003) Invited commentary: monitoring fecundity over time – if we do it, then let's do it right. Am J Epidemiol 157(2):94–97

Olsen J, Rachootin P, Schiodt AV, Damsbo N (1983) Tobacco use, alcohol consumption and infertility. Int J Epidemiol 12(2):179–184

Olsen J, Schmidt MM, Christensen K (1997) Evaluation of nature-nurture impact on reproductive health using half-siblings. Epidemiol 8(1):6–11

Olsen J, Basso O, Spinelli A, Küppers-Chinnow M, The European Study Group on Infertility and Subfecundity (1998a) Correlates of care seeking for infertility treatment in Europe. Eur J Public Health 8(1):15–20

Olsen J, Juul S, Basso O (1998b) Measuring time to pregnancy. Methodological issues to consider. Hum Reprod 13(7):1751–1753

Olsen J, Melbye M, Olsen SF, Sorensen TI, Aaby P, Andersen AM, Taxbol D, Hansen KD, Juhl M, Schow TB, Sorensen HT, Andresen J, Mortensen EL, Olesen AW, Sondergaard C (2001) The Danish National Birth Cohort – its background, structure and aim. Scand J Public Health 29(4):300–307

Olsen J, Czeizel A, Sorensen HT, Nielsen GL, de Jong van den Berg LT, Irgens LM, Olesen C, Pedersen L, Larsen H, Lie RT, de Vries CS, Bergman U (2002) How do we best detect toxic effects of drugs taken during pregnancy? A EuroMap paper. Drug Saf 25(1):21–32

Olsen J, Bonde JP, Hjøllund NH, Basso O, Ernst E (2005) Using infertile patients in epidemiologic studies on subfecundity and embryonal loss. Hum Reprod Update 11(6):607–611

O'Sullivan JB (1991) Diabetes mellitus after GDM. Diabetes 40(suppl 2):131–135

O'Sullivan JB, Mahan CM (1964) Criteria for the oral glucose tolerance test in pregnancy. Diabetes 13:278–285

Paneth N (2008) Invited commentary: the hidden population in perinatal epidemiology. Am J Epidemiol 167(7):793–796; author reply 797–798

Pedersen CB, Sun Y, Vestergaard M, Olsen J, Basso O (2007) Assessing fetal growth impairments based on family data as a tool for identifying high-risk babies. An example with neonatal mortality. BMC Pregnancy Childbirth 7:28

Pharoah PO, Adi Y (2000) Consequences of in-utero death in a twin pregnancy. Lancet 355(9215):1597–1602

Quenby S, Vince G, Farquharson R, Aplin J (2002) Opinion. Recurrent miscarriage: a defect in nature's quality control? Human Reprod 17(8):1959–1963

Rachootin P, Olsen J (1982) Prevalence and socioeconomic correlates of subfecundity and spontaneous abortion in Denmark. Int J Epidemiol 11(3):245–249

Rachootin P, Olsen J (1983) The risk of infertility and delayed conception associated with exposure in the Danish workplace. J Occup Med 25(5):394–402

Ramlau-Hansen CH, Thulstrup AM, Storgaard L, Toft G, Olsen J, Bonde JP (2007) Is prenatal exposure to tobacco smoking a cause of poor semen quality? A follow-up study. Am J Epidemiol 165(12):1372–1379

Ravelli GP, Stein ZA, Susser MW (1976) Obesity in young men after famine exposure in utero and early infancy. N Engl J Med 295(7):349–353

Reddy UM, Branum AM, Klebanoff MA (2005) Relationship of maternal body mass index and height to twinning. Obstet Gynecol 105(3):593–597

Roberts JM, Cooper DW (2001) Pathogenesis and genetics of pre-eclampsia. Lancet 357(9249):53–56

Roberts JM, Lain KY (2002) Recent insights into the pathogenesis of pre-eclampsia. Placenta 23(5):359–372

Robillard PY, Hulsey TC, Perianin J, Janky E, Miri EH, Papiernik E (1994) Association of pregnancy-induced hypertension with duration of sexual cohabitation before conception. Lancet 344(8928):973–975

Rockenbauer M, Olsen J, Czeizel AE, Pedersen L, Sorensen HT (2001) Recall bias in a case-control surveillance system on the use of medicine during pregnancy. Epidemiology 12(4): 461–466

Romero R, Mazor M, Munoz H, Gomez R, Galasso M, Sherer DM (1994) The preterm labor syndrome. Ann N Y Acad Sci 734:414–429

Romero R, Espinoza J, Kusanovic JP, Gotsch F, Hassan S, Erez O, Chaiworapongsa T, Mazor M (2006) The preterm parturition syndrome. BJOG 113(suppl 3):17–42

Rothenberg SJ, Kondrashov V, Manalo M, Jiang J, Cuellar R, Garcia M, Reynoso B, Reyes S, Diaz M, Todd AC (2002) Increases in hypertension and blood pressure during pregnancy with increased bone lead levels. Am J Epidemiol 156(12):1079–1087

Royal College of Physicians of London (1977) Smoking or health: the third report from the Royal College of Physicians of London. Pitman Medical, Kent

Sabroe S, Olsen J (1998) Perinatal correlates of specific histological types of testicular cancer in patients below 35 years of age: a case-cohort study based on midwives' records in Denmark. Int J Cancer 78(2):140–143

Salas SP (1999) What causes pre-eclampsia? Baillieres Best Pract Res Clin Obstet Gynaecol 13(1):41–57

Savitz DA (2008) Invited commentary: disaggregating preterm birth to determine etiology. Am J Epidemiol 168:990–992; discussion 993–994

Savitz DA, Sonnenfeld NL, Olshan AF (1994) Review of epidemiologic studies of paternal occupational exposure and spontaneous abortion. Am J Ind Med 25(3):361–383

Savitz DA, Chan RL, Herring AH, Howards PP, Hartmann KE (2008) Caffeine and miscarriage risk. Epidemiology 19(1):55–62

Schmidt MI, Matos MC, Reichelt AJ, Forti AC, de Lima L, Duncan BB (2000) Prevalence of gestational diabetes mellitus – do the new WHO criteria make a difference? Brazilian Gestational Diabetes Study Group. Diabet Med 17(5):376–380

Schmidt MI, Duncan BB, Reichelt AJ, Branchtein L, Matos MC, Costa e Forti A, Spichler ER, Pousada JM, Teixeira MM, Yamashita T (2001) Gestational diabetes mellitus diagnosed with a 2-h 75-g oral glucose tolerance test and adverse pregnancy outcomes. Diabetes Care 24(7):1151–1155

Sharpe RM, Skakkebaek NE (1993) Are oestrogens involved in falling sperm counts and disorders of the male reproductive tract? Lancet 341(8857):1392–1395

Shepard TH, Fantel AG, Fitzsimmons J (1989) Congenital defect rates among spontaneous abortuses: twenty years of monitoring. Teratology 39(4):325–331

Shrimpton R (2003) Preventing low birthweight and reduction of child mortality. Trans R Soc Trop Med Hyg 97(1):39–42

Sibai BM (1998) Prevention of preeclampsia: a big disappointment. Am J Obstet Gynecol 179(5):1275–1278

Skjaerven R, Gjessing HK, Bakketeig LS (2000) New standards for birth weight by gestational age using family data. Am J Obstet Gynecol 183(3):689–696

Skjaerven R, Wilcox AJ, Lie RT (2002) The interval between pregnancies and the risk of preeclampsia. N Engl J Med 346(1):33–38

Skjaerven R, Vatten LJ, Wilcox AJ, Rønning T, Irgens LM, Lie RT (2005) Recurrence of pre-eclampsia across generations: exploring fetal and maternal genetic components in a population based cohort. BMJ 331(7521):877

Smith GD, Ebrahim S (2003) ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease. 30th Thomas Francis Jr. Memorial Lecture, delivered by George Davey Smith at the University of Michigan, School of Public Health, 6 Mar 2003. Int J Epidemiol 32:1–22

Smith GD, Ebrahim S (2005) Mendelian randomization: prospects, potentials, and limitations. Int J Epidemiol 34(2):481–482

Sondergaard C, Skajaa E, Henriksen TB (2000) Fetal growth and infantile colic. Arch Dis Child Fetal Neonatal Ed 83(1):F44–F47

Sørensen HJ, Mortensen EL, Reinisch JM, Mednick SA (2009) Association between prenatal exposure to bacterial infection and risk of schizophrenia. Schizophr Bull 35(3):631–637

Stang A, Ahrens W, Bromen K, Baumgardt-Elms C, Jahn I, Stegmaier C, Krege S, Jöckel KH (2001) Undescended testis and the risk of testicular cancer, importance of source and classification of exposure information. Int J Epidemiol 30(5):1050–1056

Starr T, Levine RJ (1983) Assessing effects of occupational exposure on fertility with indirect standardization. Am J Epidemiol 118(6):897–904

Statens Serum Institut (2013) Danish National Birth Cohort. http://www.ssi.dk/English/RandD/Research%20areas/Epidemiology/DNBC.aspx. Accessed 1 May 2013

Steel AJ, Sutcliffe A (2009) Long-term health implications for children conceived by IVF/ICSI. Hum Fertil (Camb) 12(1):21–27

Storgaard L, Bonde JP, Ernst E, Andersen CY, Kyvik KO, Olsen J (2002) Effect of prenatal exposure to oestrogen on quality of semen: comparison of twins and singleton brothers. BMJ 325(7358):252–253

Storgaard L, Bonde JP, Olsen J (2006) Male reproductive disorders in humans and prenatal indicators of estrogen exposure. A review of published epidemiological studies. Reprod Toxicol 21(1):4–15

Strohsnitter WC, Noller KL, Hoover RN, Robboy SJ, Palmer JR, Titus-Ernstoff L, Kaufman RH, Adam E, Herbst AL, Hatch EE (2001) Cancer risk in men exposed in utero to diethylstilbestrol. J Natl Cancer Inst 93(7):545–551

Sun Y, Vestergaard M, Christensen J, Nahmias AJ, Olsen J (2008) Prenatal exposure to maternal infections and epilepsy in childhood: a population-based cohort study. Pediatrics 121(5): e1100–e1107

Sun Y, Strandberg-Larsen K, Vestergaard M, Christensen J, Nybo Andersen AM, Grønbaek M, Olsen J (2009) Binge drinking during pregnancy and risk of seizures in childhood: a study based on the Danish National Birth Cohort. Am J Epidemiol 169(3):313–322

Susser M, Stein Z (1994) Timing in prenatal nutrition: a reprise of the Dutch Famine Study. Nutr Rev 52(3):84–94

Swan SH, Shaw GM, Schulman J (1992) Reporting and selection bias in case-control studies of congenital malformations. Epidemiology 3(4):356–363

Tentoni S, Astolfi P, De Pasquale A, Zonta LA (2004) Birthweight by gestational age in preterm babies according to a Gaussian mixture model. BJOG 111(1):31–37

Tong S, Caddy D, Short RV (1997) Use of dizygotic to monozygotic twinning ratio as a measure of fertility. Lancet 349(9055):843–845

Trabert B, Zugna D, Richiardi L, McGlynn KA, Akre O (2013) Congenital malformations and testicular germ cell tumors. Int J Cancer. [Epub ahead of print] PubMed PMID: 23580254

Trichopoulos D (1990) Hypothesis: does breast cancer originate in utero? Lancet 335(8695): 939–940

Trogstad LI, Eskild A, Magnus P, Samuelsen SO, Nesheim BI (2001) Changing paternity and time since last pregnancy; the impact on pre-eclampsia risk. A study of 547 238 women with and without previous pre-eclampsia. Int J Epidemiol 30(6):1317–1322

Trupin LS, Simon LP, Eskenazi B (1996) Change in paternity: a risk factor for preeclampsia in multiparas. Epidemiology 7(3):240–244

VanderWeele TJ, Hernández-Diaz S (2011) Is there a direct effect of pre-eclampsia on cerebral palsy not through preterm birth? Paediatr Perinat Epidemiol 25(2):111–115

VanderWeele TJ, Mumford SL, Schisterman EF (2012) Conditioning on intermediates in perinatal epidemiology. Epidemiology 23(1):1–9. Erratum in: Epidemiology 23(3):507

Verdoux H, Geddes JR, Takei N, Lawrie SM, Bovet P, Eagles JM, Heun R, McCreadie RG, McNeil TF, O'Callaghan E, Stober G, Willinger MU, Wright P, Murray RM (1997) Obstetric complications and age at onset in schizophrenia: an international collaborative meta-analysis of individual patient data. Am J Psychiatry 154(9):1220–1227

Vestergaard M, Basso O, Henriksen TB, Oestergaard JR, Olsen J (2002) Risk factors for febrile convulsions. Epidemiology 13(3):282–287

Villar J, Belizan JM (2000) Same nutrient, different hypotheses: disparities in trials of calcium supplementation during pregnancy. Am J Clin Nutr 71(5 suppl):1375S–1379S

von Dadelszen P, Ornstein MP, Bull SB, Logan AG, Koren G, Magee LA (2000) Fall in mean arterial pressure and fetal growth restriction in pregnancy hypertension: a meta-analysis. Lancet 355(9198):87–92

Watier L, Richardson S, Hémon D (1997) Accounting for pregnancy dependence in epidemiologic studies of reproductive outcomes. Epidemiology 8(6):629–636

Weinberg CR (1993) Toward a clearer definition of confounding. Am J Epidemiol 137(1):1–8

Weinberg CR, Wilcox AJ (2008) Methodologic issues in reproductive epidemiology. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 3rd edn. Lippincott Williams & Wilkins, Philadelphia

Weinberg CR, Baird DD, Rowland AS (1993) Pitfalls inherent in retrospective time-to-event studies: the example of time to pregnancy. Stat Med 12(9):867–879

Weinberg CR, Baird DD, Wilcox AJ (1994a) Sources of bias in studies of time to pregnancy. Stat Med 13(5–7):671–681

Weinberg CR, Baird DD, Wilcox AJ (1994b) Bias in retrospective studies of spontaneous abortion based on the outcome of the most recent pregnancy. Ann N Y Acad Sci 18(709):280–286

Weinberg CR, Wilcox AJ, Lie RT (1998) A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. Am J Hum Genet 62(4):969–978

Wen SW, Demissie K, Liu S, Marcoux S, Kramer MS (2000) Placenta praevia and male sex at birth: results from a population-based study. Paediatr Perinat Epidemiol 14(4):300–304

Weng X, Odouli R, Li DK (2008) Maternal caffeine consumption during pregnancy and the risk of miscarriage: a prospective cohort study. Am J Obstet Gynecol 198(3):279.e1–e8

Whitacre CC (2001) Sex differences in autoimmune disease. Nat Immunol 2(9):777–780

Whitacre CC, Reingold SC, O'Looney PA (1999) A gender gap in autoimmunity. Science 283(5406):1277–1278

Wilcox AJ (2001) On the importance – and the unimportance – of birthweight. Int J Epidemiol 30(6):1233–1241

Wilcox AJ, Russell IT (1986) Birthweight and perinatal mortality: III. Towards a new method of analysis. Int J Epidemiol 15(2):188–196

Wilcox AJ, Weinberg CR, O'Connor JF, Baird DD, Schlatterer JP, Canfield RE, Armstrong EG, Nisula BC (1988) Incidence of early loss of pregnancy. N Engl J Med 319(4):189–194

Wilcox AJ, Baird DD, Weinberg CR, Hornsby PP, Herbst AL (1995) Fertility in men exposed prenatally to diethylstilbestrol. N Engl J Med 332(21):1411–1416

Wilcox AJ, Weinberg CR, Basso O (2011) On the pitfalls of adjusting for gestational age at birth. Am J Epidemiol 174(9):1062–1068

Winbo IG, Serenius FH, Dahlquist GG, Kallen BA (1997) A computer-based method for cause of death classification in stillbirths and neonatal deaths. Int J Epidemiol 26(6):1298–1306

Xiao J, Buka SL, Cannon TD, Suzuku Y, Viscidi RP, Torrey EF, Yolken RH (2009) Serological pattern consistent with infection with type I Toxoplasma gondii in mothers and risk of psychosis among adult offspring. Microbes Infect 11(13):1011–1018

Xiong X, Demianczuk NN, Saunders LD, Wang FL, Fraser WD (2002) Impact of preeclampsia and gestational hypertension on birth weight by gestational age. Am J Epidemiol 155(3): 203–209

Yerushalmy J (1964) Mother's cigarette smoking and survival of infant. Am J Obstet Gynecol 88:505–518

Yudkin PL, Wood L, Redman CW (1987) Risk of unexplained stillbirth at different gestational ages. Lancet 1(8543):1192–1194

Zhang J, Cai WW (1991) Severe vomiting during pregnancy: antenatal correlates and fetal outcomes. Epidemiology 2(6):454–457

Zhu S (2009) Psychosis may be associated with toxoplasmosis. Med Hypotheses 73(5): 799–801

Zhu JL, Basso O, Obel C, Bille C, Olsen J (2006) Infertility, infertility treatment, and congenital malformations: Danish National Birth Cohort. BMJ 333(7570):679

Zhu JL, Obel C, Hammer Bech B, Olsen J, Basso O (2007a) Infertility, infertility treatment, and fetal growth restriction. Obstet Gynecol 110(6):1326–1334

Zhu JL, Basso O, Obel C, Christensen K, Olsen J (2007b) Infertility, infertility treatment and twinning: Danish National Birth Cohort. Hum Reprod 22(4):1086–1090

Zhu JL, Basso O, Obel C, Hvidtjørn D, Olsen J (2009) Infertility, infertility treatment and psychomotor development: Danish National Birth Cohort. Paediatr Perinat Epidemiol 23(2): 98–106

Further Reading

Adams M, Alexander GR, Kirby RS, Wingate MS (eds) (2009) Perinatal epidemiology for public health practice. Springer, New York

Bracken MB (ed) (1984) Perinatal epidemiology. Oxford University Press, New York

Kallen B (1988) Epidemiology of human reproduction. CRC, Boca Raton

Keith LG, Papiernik E, Keith DM, Luke B (eds) (1995) Multiple pregnancy: epidemiology, gestation and perinatal outcome, 1st edn. Parthenon Publishing Group, New York

Kiely M (ed) (1991) Reproductive and perinatal epidemiology. CRC, Boca Raton

McDowall ME (1985) Occupational reproductive epidemiology: the use of routinely collected statistics in England and Wales, 1980–82. Her Majesty's Stationery Office, London

Murray CJL, Lopez AD (eds) (1988) Health dimensions of sex and reproduction. Global burden of disease and injury series, vol III. WHO, Geneva

Wilcox AJ (2010) Fertility and pregnancy: an epidemiologic perspective, 1st edn. Oxford University Press, New York

Molecular Epidemiology 46

Salina M. Torres, Esther Erdei, Marianne Berwick,
Giuseppe Matullo, and Paolo Vineis

Contents

S.M. Torres (✉) • E. Erdei • M. Berwick
Division of Epidemiology and Biostatistics, University of New Mexico, Albuquerque, NM, USA

G. Matullo
Genomic Variation in Human Population and Complex Diseases, Human Genetics Foundation (HuGeF), Turin, Italy

Department of Medical Science, University of Turin, Turin, Italy

P. Vineis
Chair of Environmental Epidemiology, School of Public Health, Imperial College London, London, UK

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_28,

46.1 Introduction

Molecular epidemiology can be defined as the application of the techniques of molecular biology to the study of the epidemiology of disease in human populations. Molecular investigations, as we will see, have several aims and can contribute to the elucidation of disease etiology. In molecular epidemiology, the study of the determinants of disease focuses on causative, protective, or predisposing factors (including infectious agents and a variety of environmental exposures such as chemical or physical agents and lifestyle habits) and host characteristics such as genetic susceptibility. These studies are performed at the molecular level using the techniques of molecular biology.

Markers used in molecular epidemiology are usually divided into the three categories: *markers of exposure, markers of biological effect, and markers of susceptibility* (Vineis and Perera 2007). The underlying concept is that there is *continuity* between exposure to a toxic agent (such as carcinogens), metabolism (activation or deactivation), adduction to proteins or deoxyribonucleic acid (DNA) (i.e., formation of links by active metabolites), DNA alterations (such as mutations or chromosome damage), onset of disease, and finally, progression of disease. These concepts are schematically represented in Fig. 46.1, from Vineis and Perera (2007).

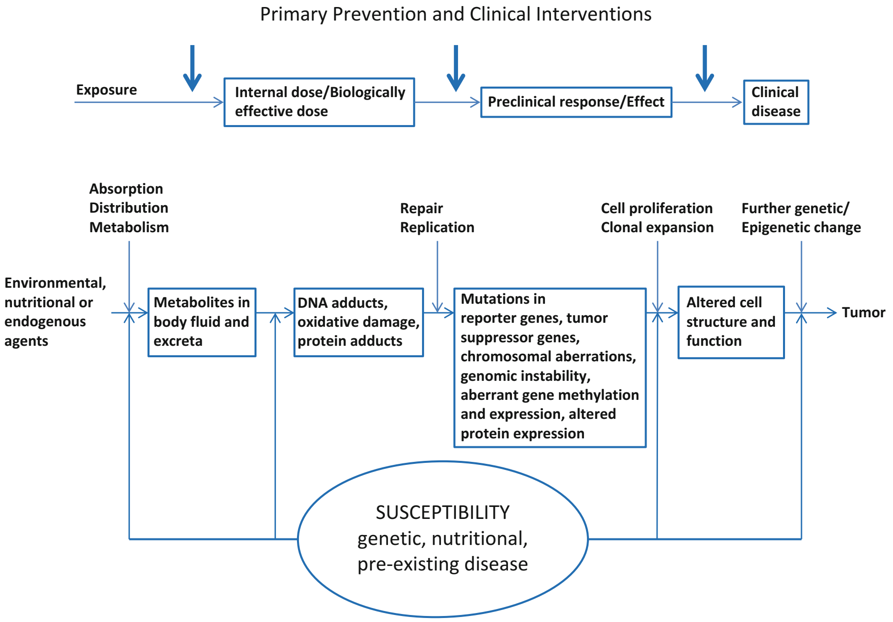

Primary Prevention and Clinical Interventions
Exposure
Internal dose/Biologically effective dose
Preclinical response/Effect
Clinical disease
Absorption
Distribution
Metabolism
Repair
Replication
Cell proliferation
Clonal expansion
Further genetic/
Epigenetic change
Environmental, nutritional or endogenous agents
Metabolites in body fluid and excreta
DNA adducts, oxidative damage, protein adducts
Mutations in reporter genes, tumor suppressor genes, chromosomal aberrations, genomic instability, aberrant gene methylation and expression, altered protein expression
Altered cell structure and function
Tumor
SUSCEPTIBILITY
genetic, nutritional,
pre-existing disease

Fig. 46.1 Updated model for molecular epidemiology – large arrows indicate points for intervention (Adapted after Vineis and Perera (2007))

In contrast to classical epidemiology, molecular and genetic epidemiology enhances the understanding of the pathogenesis of disease by identifying specific molecular pathways and specific molecules and genes that influence the risk of developing disease. For example, genetic markers rather than surrogate information (i.e., positive family history) might more precisely characterize host susceptibility. Molecular epidemiology can also improve the validity of and reduce bias in the assessment of environmental exposures. It can help to evaluate markers at the subclinical level, very early during disease onset. Molecular epidemiology provides tools to ultimately reduce or clarify heterogeneity within a disease, such as the development of breast cancer subtypes (i.e., basal, luminal A, luminal B, normal breast-like, and ERBB2+ (Kapp et al. 2006)). However, most of the expectations stemming from molecular epidemiology need to be carefully verified in well-designed studies.

Not all molecular biology markers are suitable for molecular epidemiology studies as many are extremely labor intensive and/or expensive to use. It is critical to test all laboratory techniques, but even more so those used in molecular epidemiology studies, for their validity, sensitivity and specificity, and variability within and between laboratories before using them in an epidemiological setting. Access to appropriate biological specimens, ethical issues, and costs are also important considerations in molecular epidemiology.

This chapter includes an introduction to the main features of molecular epidemiology and sections addressing biomarkers of exposure and effect, host susceptibility, issues in using biomarkers in epidemiology, common study designs, risk assessment, and the use of biorepositories. The concept of Mendelian randomization will be also introduced describing the use of genetic markers to better address possible bias and confounding in exposure assessment. This chapter is meant to present an overview; the authors refer the reader to the book edited by Wild et al. (2008) for more detailed explanations of the topics discussed here.

Critical issues in molecular epidemiology include those relevant to any epidemiological study but are particularly important in molecular epidemiology: an appropriate study design, careful attention to sources of bias and confounding, and the development of markers that can be applied on a population scale. The study design is particularly important because past research that applied laboratory methods to human populations was often based on “convenience samples” – that is, groups of patients recruited in the most comfortable efficient way without a proper design – that were frequently affected by bias.

Study design is thoroughly covered in several chapters of Part I (Concepts and Designs in Epidemiology) of this handbook. Within this chapter, we will focus on specificities of molecular epidemiology. Important issues associated with epidemiological studies, such as confounding and interaction (see also chapter ▸Confounding and Interaction of this handbook) and measurement error (see also chapter ▸Measurement Error of this handbook), are dealt with in detail in this book.

46.2 Classes of Biomarkers

46.2.1 Biomarkers of Exposure

The International Programme on Chemical Safety (IPCS 1999) defines human exposure as any contact with chemical, physical and biological agents. Molecular epidemiology studies seek to establish causality and biological plausibility for associations between exposure and disease (Wild 2009). Exposure assessment is one aspect of this continuum, and biomarkers of exposure may contribute to this assessment by providing more precise information related to the exposure-disease association. Chapter ▸Exposure Assessment of this handbook gives a thorough explanation of the definition, data, assessment, and the role of measurement error in exposure assessment. The role of exposure assessment in molecular epidemiology is thus only briefly described here.

Molecular epidemiology utilizes biomarkers to refine exposure assessment in the face of the complexity of distinguishing between the contribution of environmental and individual factors to disease etiology (Wild 2009). Validated biomarkers have the potential to provide measurements at the individual level that are relevant to events on, or related to, the causal pathway and information applicable to the biological plausibility of an exposure-disease association. For example, in the field of virology, antibodies have long been used to identify the type of virus that a person has been infected with. Further, the measurement of accumulation of chemical agents or metals in biological tissues like arsenic in hair or mercury in fingernails and toenails provides direct measurements of an exposure at the individual level.

Using sensitive laboratory technologies, biomarkers of exposure can detect low levels of exposure. Although these biomarkers are usually designed to measure exposure to environmental toxicants, they can also identify levels of ingested dietary components, bacterial and viral infections, as well as serve as endpoints in the determination of the success of intervention strategies.

Biomarkers of exposure are classified depending on what they measure, either an internal dose such as serum vitamin D or a biologically effective dose (i.e., a dose that causes DNA damage). Internal dose biomarkers measure the presence of environmental chemicals and their metabolites in human tissues, excretions, and/or exhaled air. Additionally, measurements can be made using dietary biomarkers either as “recovery” biomarkers, such as measurements of sucrose and fructose in 24-hour urine samples to directly assess sugar consumption, or “concentration” biomarkers, such as serum carotenoids, which are related to dietary intake but only indirectly because their levels are the result of complex metabolic processes. Exposure biomarkers can also be applicable in intervention strategies aimed at primary prevention or mechanism-based approaches like chemoprevention, that is, biomarkers can be used as endpoints to measure the success of the intervention. The utility of such biomarkers is limited to the availability of detectable levels of the compound, its metabolites, and/or antibodies and proteins, discussed below.

Although biomarkers of exposure are often described as a separate topic from biomarkers of effect, many actually overlap and can be utilized as both. This overlap can be attributed in part to the fact that the biomarkers provide information related to both endpoints. In addition, the same tissue(s) and/or cells may be used, for example, lymphocytes can be a surrogate for exposure and also a target for the exposures' effect.

46.2.2 Biomarkers of Effect

Biomarkers of effect measure the interaction between an agent and/or its metabolites and target cell(s) or molecule(s); they are defined as "measureable changes in the organism" (IPCS 1999). This definition also includes markers of effect that can indicate a preclinical response not always detectable using conventional clinical diagnostic tools. Early effects within an individual may also be used as informative markers of disease risk. Several molecular-based assays have been developed to identify cellular response(s) stimulated by such exposures. Some of these assays are listed in Table 46.1, and their utility as a marker of both exposure and effect is discussed.

Biomarkers of a biologically effective dose measure the amount of the agent reaching a critical cellular target, such as DNA adducts or the amount of a chemical agent bound to a cellular receptor. Other markers of effect measure the damage that is induced by the agent. For example, in the case of a biologically effective UV dose, UV-induced DNA damage is measured, and the type of damage can be further characterized by the wave length, that is, DNA damage from UVA exposure induces base excision repair, while UVB exposure induces nucleotide excision repair. Biomarkers of DNA damage include mutations, adducts, chromosomal aberrations, micronuclei, DNA strand breaks, and sister chromatid exchanges, among others. Additional examples of biomarkers can be found in Box 46.1 which provides a list of measures used in molecular epidemiology that is organized by biological tissue.

46.3 Biomarker Selection

There are several issues to consider when identifying candidate biomarkers. In the case of molecular epidemiology studies that evaluate biomarkers, the prevalence of the biomarker of interest in the population being studied should be known. The ability of the biomarker to represent the agent of interest and the specificity and sensitivity of the biomarker to detect low levels of exposure are additional factors to take into consideration when selecting appropriate biomarkers of exposure. Additionally, the validity of the biomarker, defined as the lack of systematic measurement error when compared to a standard reference, must be established. Epidemiological studies utilizing biomarkers must take into account that environmental exposures will vary qualitatively and quantitatively over time.

Table 46.1 Examples of laboratory measurements that can be utilized as biomarkers of exposure and effect

Marker	Description	Biomarker of exposure	Biomarker of effect
Measurement of genetic alterations			
DNA damage (Olive and Banath 2006)	Comet assay – lymphocytes are treated and run out on a gel; DNA from the cell “migrates” to form a “tail.” The length of the tail is proportional to the amount of DNA breakage	Measures DNA damage as a result of exposure, without time for repair	Cells are subjected to a damaging agent and allowed time to repair. Then their DNA repair capacity is measured by the length of the comet tail
Chromosome aberrations (CA) (Bonassi and Au 2002)	Detection of altered chromosome formations in blood lymphocytes	CAs are specific to the agent of exposure, allow for the ability to distinguish the type of exposure	Can predict survival of cells with certain CAs, thought to have the ability to predict cancer risk
Micronuclei (MN) (Neri et al. 2003; Bonassi and Au 2002)	Detects micronuclei after the cells have progressed past the first cell cycle, MN contain either chromosomal fragments or whole chromosomes	Useful for biomonitoring studies of exposure, including low-dose environmental exposures	Can distinguish the cause of MN formation: chromosome breakage or aneuploidy
Sister chromatid exchange (SCE) (Albertini et al. 2000)	Formation of SCEs correlates to the effectiveness of recombinational DNA repair and induction of point mutations, gene amplification, and cytotoxicity	Alterations in SCE levels may be correlated to exposure levels	Indicator of DNA damage, can evaluate altered cell viability induced by genetic defects or exogenous exposures

Hypoxanthine phosphoribosyltransferase (HPRT) (Bonassi and Au 2002)	Gene mutation assay, detects somatic mutation in this hemizygous gene in the X chromosome in post-thymic lymphocytes	Used for detection of mutation(s) in exposed populations, can evaluate the mutagenic potential of various compounds (Perera et al. 2002)	Mutation frequency has been found to correlate with the presence of DNA adducts, has also been found to correlate with exposure level (Ammenheuser et al. 2001)
Epigenetic modifications	Changes in the genome that do not involve alterations in the nucleotide sequence can be measured using various techniques	CpG-promoter sequence methylation, histone modification, and chromatin remodeling	Alterations in the balance of methylation are related to gene transcriptional silencing
Glycophorin A (GPA) gene mutations	Detects damage in bone marrow cells, damage is measured as the loss of alleles	Mutant frequency can be correlated to exposure	GPA levels can be correlated to metabolite levels
DNA adducts	Measure of DNA damage	Levels of DNA adducts can be correlated with exposure levels	Unrepaired DNA adduct levels reflect DNA repair capacity
Measurement of other biological alterations			
Transcriptomics	Detection of altered gene expression	Alterations in gene expression vary depending on the type of exposure	Changes in gene expression profiles in response to chronic exposure
Proteomics	High-throughput broad screen of proteins using mass spectrometry	Albumin and hemoglobin adducts: changes in protein structure as a consequence of exposure	Changes in protein expression as a consequence of exposure (i.e., altered growth factors and cytokines)
Metabolomics	Detection of metabolites, defining metabolic profiles to predict onset of disease	Accumulation of metabolites of exposure agents	Changes in the metabolic capacity as a consequence of exposure or disease

Box 46.1. Common biological samples used for biomarker analysis in molecular epidemiological studies, listed according to the source of the material. It should be noted that this list will continually change over time

1. DNA
   - (a) Genomic DNA
     - Single nucleotide polymorphisms (SNPs) (>1% prevalence)
     - Epigenetic modifications (i.e., methylation, histone modification, and chromatin remodeling)
     - Adducts
     - Copy-number variations (CNVs)
   - (b) Tumor DNA
     - Somatic mutations (>1% prevalence)
     - Copy-number variations (CNVs)
   - (c) Mitochondrial DNA
     - Insertions and deletions

2. RNA
   - Viral genotyping
   - Microarray chips for gene expression profiles

3. Whole cells
   - (a) Lymphocytes
     - Incorporation of damaged plasmid (host-cell reactivation assay (HCRA))
     - Comet assay
     - Hypoxanthine phosphoribosyltransferase assay in peripheral T lymphocytes (HPRT) assay
   - (b) Chromosomes
     - Cytogenetic assays to assess mutagen sensitivity (mutagen sensitivity assay (MSA))
     - Chromosome breaks and deletions (chromosomal aberrations CAs)
     - Sister chromatid exchanges (SCEs)
   - (c) Shed cells – for extraction of DNA, RNA, measurement of DNA damage, DNA repair, adducts, micronuclei, and metabolomics in:
     - Exfoliated bladder cells
     - Oral buccal cells
     - Broncholavage
   - (d) Glycophorin A (GPA) gene mutations in erythrocytes

4. Plasma/serum
   - Extraction of tumor and genomic DNA from plasma and serum
   - Measurement of indicators of immune response (antibodies, cytokines, T lymphocytes, etc.)
   - Proteomics – albumin adducts

Metabolomics
Measurement of biochemicals, i.e., vitamin D and folate

5. Red blood cells
   Hemoglobin, hemoglobin adducts, and biochemical content, i.e., folate

6. Urine
   Urinary metabolites – biochemical assays, i.e., cotinine
   Exfoliated bladder cells – for DNA, see above

7. Hair
   Biological accumulation, i.e., arsenic and cotinine

8. Fingernails, toenails
   Biological accumulation, i.e., arsenic, and mercury

Another consideration is that some biomarkers decay over their lifetime, so that the half-life of biomarkers must be considered when choosing an appropriate design for a molecular epidemiological study. Most biomarkers of exposure are transient with a relatively short biologically relevant half-life. In "traditional" epidemiology, a case-control study design is particularly useful when the disease being studied is rare and the exposure hypothesized is frequent and easily identifiable. Studies aimed at investigating cancer and other chronic diseases are usually focused on events that have taken place many years before the disease onset and often involve chronic exposures. In the molecular epidemiology of chronic diseases, the case-control approach is limited if the potential biomarker has a short half-life, that is, it refers to an acute exposure that took place a short time before disease onset. In the case of adducts, those measured in the DNA of lymphocytes have a half-life of months, and those measured in hemoglobin have a half-life of weeks. The value of DNA adducts as biomarkers is discussed in detail below. These issues demonstrate the utility of prospective studies for molecular epidemiological investigations; however, such studies are often limited in the sense that they are based on the use of a "one-time" biological sample that is not necessarily representative of the usual exposure or of changing exposures (thus, introducing misclassification, see chapter ►Misclassification of this handbook).

Biomarker validation encompasses several issues that require consideration when assessing the utility of a biomarker for a molecular epidemiological study. Analytical validity, *c*linical validity, *c*linical utility, and ethical, legal, and social implications and safeguards are often referred to as the *ACCE* evaluation of a biomarker (Vineis and Gallo 2007). The analytical validity component of this evaluation refers to the ability of a test to accurately and reliably measure the marker/genotype of interest which includes its sensitivity and specificity. The ability of a genetic test to detect or predict the phenotype is the focus of the clinical validity or the positive predictive value. The clinical utility component of this evaluation considers the risks and benefits associated with the incorporation of the test into routine clinical practice. Lastly, the ethical, legal, social implications and safeguards are other components that require consideration when evaluating the utility of a biomarker. For a more detailed description of the principles of biomarker validation, the reader is referred to the recent book edited by Vineis and Gallo (2007).

Betsou and colleagues (2009) recommended several types of assessments to evaluate the vulnerability of a biomarker to pre-analytical variation; these assessments can be utilized for determination and evaluation of candidate biomarkers to ensure that association with clinical endpoints is not due to uncontrolled pre-analytical variation. Biospecimen research focuses on artifactual variation as opposed to true variation that may compromise results; if such variation occurs before the laboratory analysis, it is referred to as pre-analytical variation (see the Sect. 46.8 of this chapter). Clearly, if serum is not handled properly, the characteristics can change quickly (e.g., vitamin C is light sensitive). Empirical biospecimen research is recommended to assess the impact of pre-analytical variation on biomarkers. Proteomic markers, for example, are particularly sensitive to the need for consistent handling, or the results are meaningless. Other sources of pre-analytical variation include the time of specimen collection, fasting conditions, the position of the patient during collection, the patient's diet, or other life habits, all of which must be considered when selecting an appropriate biomarker (Betsou et al. 2009).

The selection of an inaccurate biomarker may partly result from publication bias toward false-positive associations. There are likely many biomarkers tested but never published due to poor results. Publication bias may be a result of time-consuming and costly assays, such that manuscripts reporting positive results are more likely to be published than negative findings, although the results may have been obtained by chance due to a small sample size.

46.4 Types of Biomarkers: Examples

Technological advances continue to produce innovative molecular biological techniques. The following section describes some common molecular-based assays used in molecular epidemiological studies. Some of the biomarkers are relatively old and well established; others are novel. The descriptions of the assays below include recent validation efforts and issues to consider when employing these techniques. Table 46.1 contains examples of some of the laboratory methods described in this section.

46.4.1 DNA Adducts

Chemically induced DNA adducts are an indication of the mutagenicity of chemical agents as well as of an elevated cancer risk (Angerer et al. 2007). Studies assessing exposure to air pollution have utilized DNA damage as an endpoint (Vineis and Husgafvel-Pursiainen 2005). In particular, several studies have found an association between external measures of exposure to air pollution and increased levels of DNA adducts, with an apparent leveling off of the dose-response relationship (reviewed in Vineis and Husgafvel-Pursianinen (2005)).

The use of DNA adducts as biomarkers requires consideration of several issues: (i) reproducibility of adduct levels measured with different methods, (ii) correlation between levels of the same type of DNA adduct in normal versus surrogate tissue and/or normal versus tumor tissue, and (iii) correlation between different biomarkers of exposure. Although DNA damage is a reflection of exposure, the resulting detectable damage is influenced by inherent and acquired susceptibilities (Vineis and Husgafvel-Pursianinen 2005; Gyorffy et al. 2008).

As an example, DNA adducts and urinary metabolites have been examined to measure the correlation between polycyclic aromatic hydrocarbons (adduct) and prenatal environmental tobacco smoke (Perera et al. 2007). The study was designed to measure prenatal exposure and its effect on postnatal development.

46.4.2 Epigenetic Markers

Cancer epidemiology has traditionally focused on evaluating exposure to carcinogens by monitoring genomic biomarkers of DNA damage. Recently the focus has shifted to include biomarkers of exposure that represent changes in epigenetic mechanisms. Epigenetic modifications are changes in the genome that do not involve alterations in the nucleotide sequences and that can pass from somatic cells to their daughters. These modifications have been thought to explain the disconnection between genotype and gene expression (Vineis and Perera 2007). In particular, methylation of CpG promoter sequences (CpG sequences are areas in the genome, cytosine-guanine nucleotides separated by a phosphate, that are particularly prone to methylation), histone modification, chromatin remodeling, microRNAs, and receptor binding have been identified in cancer and used as epigenetic biomarkers of exposure and of disease susceptibility (Wild 2009).

DNA methylation is thought to play a role in suppressing gene expression subsequently affecting protein synthesis. Alterations in the balance of intrinsic methylation, such as hypermethylation of promoter regions, are related to gene transcriptional silencing, which in the case of human cancer, is a common mechanism for the inactivation of tumor suppressor genes (Robertson and Wolffe 2000). This example of gene silencing, an epigenetic alteration, demonstrates the relevance of epigenetic changes as biomarkers. Utilizing epigenetic changes as biomarkers of exposure has become a promising tool since some of the epigenetic events are reversible and can provide points of intervention for cancer prevention and therapeutic agents. Conversely, some of the epigenetic components are known to be inherited and in the context of molecular epidemiological investigations could provide an opportunity to evaluate disease susceptibility. Additional research is necessary to determine the mechanism underlying exposure-induced as well as inherent alterations in epigenetic processes which in turn alter cellular function and cancer risk (Wild 2009).

46.4.3 RNA-Based Markers

One example of an important use of RNA-based platforms to investigate a disease is that by Kang et al. (2010) in which they utilized microarray analysis to determine whether gene expression profiling could be used to improve risk classification and outcome prediction in pediatric acute lymphoblastic leukemia patients at high risk for relapse. In this study, a 38-gene expression classifier predictive of relapse-free survival which significantly enhanced outcome prediction and risk classification was identified using post treatment diagnostic samples in blood.

46.4.4 "Omic" Studies

Transcriptomics, proteomics, and metabolomics express recent advances in high-throughput technologies that have the potential to identify previously unobservable responses to exposure, assuming that the exposure of interest is reflected in the body by altered levels of specific mRNA, proteins, or metabolites, respectively. These techniques could reveal additional specific targets for exposure assessment. To mention two examples, a small pilot study evaluating prenatal arsenic exposure (Fry et al. 2007) and another one assessing tobacco smoke exposure (Spira et al. 2004) have both suggested that changes in gene expression, measured by transcriptomic technology, are related to exposure and vary depending on the type of exposure.

Proteomics The analysis of total protein output encoded by the genome is referred to as proteomics. To date, this technology has mainly been utilized to distinguish proteomic profiles between normal, benign, and disease states (Vineis and Perera 2007).

Recently, global proteomic approaches have been developed to investigate environmental exposure and subsequent biological alterations using mass spectrometry (Colquhoun et al. 2009). This high-throughput approach based on a broad screen of proteins can be utilized to identify changes in protein expression in response to exposure. Colquhoun and colleagues (2009) utilized cord blood to characterize the fetal serum proteome of infants exposed in utero to maternal cigarette smoke. In their validation study, they were able to identify candidate biomarkers that were biologically related to fetal chemical exposure where expression levels of these proteins were different from expression levels of unexposed infants. The utility of blood serum and plasma for this approach makes it applicable to large molecular epidemiological studies. Although this proteome-wide approach is still under development and requires further validation, the potential exists for further adaptation of proteomics to the epidemiological analysis of chemical exposures (Colquhoun et al. 2009).

Metabolomics Metabolomics provides a method for defining metabolic profiles of an individual that can be used to predict the onset of common diseases, characterization of certain disease states and metabolic disorders (Vineis and Perera 2007). Specifically, metabolomics is the "systematic study of the unique chemical fingerprints that specific cellular processes leave behind," the study of their small-molecule metabolite profiles using advanced technologies like nuclear magnetic resonance spectroscopy and liquid and gas chromatography (Psychogios et al. 2011; Davis 2005).

Validation of this technology has only been reported in a few published studies, to date. However, it is becoming evident that metabolomics can have broad applications to exposure assessment by investigating metabolites from many different agents.

A review of recent studies utilizing transcriptomics, proteomics, and/or metabolomics can be found in the 2009 review paper by Wild. The initial findings warrant further investigations to validate these technologies as well as to determine the sensitivity and specificity of potential biomarkers of exposure measured.

46.4.5 Cell-Based Assays

The Comet Assay The "Comet" assay was developed in the late 1980s/early 1990s and uses only a few lymphocytes. The lymphocytes have to be frozen at a very low temperature to assure their viability. The lymphocytes are then treated and run out on a gel that is spread on a glass slide. DNA from the cell "migrates" to form a "tail." If DNA is "broken" (i.e., single-strand breaks or double-strand breaks), then the length of the tail is proportional to the amount of breakage (Olive and Banath 2006). This assay is amenable to measurement of DNA single-strand breaks, cross-links, base damage, and apoptotic nuclei. Cells are also subjected to damaging agents, allowed to repair, and then placed on the gel on the glass slides. In these cases, this assay is utilized as a biomarker of effect to measure DNA repair "capacity" by the length of the comet tail (Olive and Banath 2006).

The Comet assay is frequently used in human and animal studies evaluating environmental toxicant-induced and/or experimental DNA damage; the assay first was described as an exposure biomarker. The popularity of this assay exponentially increased when it was shown that – based on its high sensitivity and specificity – this method can also enable researchers to measure increased risks for different health outcomes. Extensive validation efforts have been undertaken to achieve standardization of the method and reliable application of the Comet assay by the European Standards Committee on Oxidative DNA Damage (ESCODD 2002; Collins et al. 2004).

The Mutagen Sensitivity Assay The bleomycin-induced mutagen sensitivity assay (MSA) is an in vitro measure of DNA repair capacity developed by Hsu et al. (1989). This assay is an indirect measure of both DNA damage and DNA repair expressed as chromosome "breaks per cell" (b/c) in short-term cultured lymphocytes. It is a relatively simple test in which a higher number of bleomycin-induced chromatid breaks indicate higher "mutagen sensitivity" and lower DNA repair. For example, if the number of “breaks per cell” is greater than the median of the controls, usually around 0.8, the examined subject is considered “mutagen sensitive” (Hsu et al. 1989). Mutagen sensitivity is regarded as a genetic susceptibility phenotype for various cancers; in this assay, chromosomes in metaphase are evaluated for breakage from a specific mutagen. The response to the mutagen likely involves a number of genes from different DNA repair pathways (Cloos et al. 2006) although the specific genes accounting for the response are still unknown (Caporaso 1999).

The MSA has been used to investigate the etiology of brain cancer (Bondy et al. 1996, 2001; El-Zein et al. 2001) as well as survival with brain cancer (Sigurdson et al. 1998). This assay has also been found to distinguish between individuals with cancer who will develop second malignancies and those who will not (Cloos et al. 1996, 1999, 2000), as well as those with a family history of cancer and those without (Spitz et al. 1994; Ankathil et al. 1999; Berwick et al. 2001; Zhu et al. 2002; Wu et al. 2002; Zheng et al. 2003).

One concern about phenotypic assays, such as the mutagen sensitivity assay, has been the typical wide variation in results that may depend on the timing of the assay (within subject variation), the person performing the assay (between observer variation), and the laboratory where the assay has been performed (interlaboratory variation). However, Erdei et al. (2006) have demonstrated excellent interobserver and interlaboratory correlations. Such studies have not been performed with other functional DNA repair capacity assays.

Micronucleus Assay The micronucleus (MN) assay, as introduced in Table 46.1, which detects MN, extracellular bodies, after the cells have progressed past the first cell cycle, has the ability to distinguish chromosome breaks from aneuploidy (abnormal number of chromosomes) and can detect chromosome loss. Since MN are formed from acentric chromosomal fragments or chromosomes that are not included in either daughter nuclei, they are classified depending on whether they contain chromosomal fragments or whole chromosomes (Albertini et al. 2000). This assay lends itself to molecular epidemiological studies because of the relative ease of scoring, limited costs and personnel requirements, and the precision that scoring larger numbers of cells provides. The MN assay can be performed in peripheral blood lymphocytes, epithelial cells, erythrocytes, alveolar macrophages, and fibroblasts (Neri et al. 2003). Since this assay requires that the cells under investigation survive at least one round of nuclear division, some of the damaged cells are lost before the analysis begins, and the survivability of the damaged cells detected is not known with this assay.

Neri and colleagues (2003) conducted a review of published studies that evaluated the occurrence of MN and the impact of genotoxic exposures on MN frequency in children and adolescents. Their review revealed that this cytogenetic assay is a useful, sensitive tool available for biomonitoring studies of children including those with low-dose exposures to environmental agents. The confounding effects of age, sex, and chronic and infectious diseases on MN levels were evaluated in these studies; the only variable whose influence was irrelevant to MN frequency was sex.

In 1997, the international database collaboration known as the HUman MicroNucleus (HUMN) project was initiated using a network of laboratories to compare various protocols used to determine whether and to what extent variations might affect baseline MN frequencies (Bonassi et al. 2001). The idea was that HUMN provided (i) baseline MN frequency data from various laboratories that can be evaluated for the influence of key variables (e.g., age, sex, and exposures) on MN frequencies, (ii) analysis of variation and reproducibility in laboratory protocols and methods, and (iii) a large database of individual data that permitted a prospective cohort study linking MN frequency data with disease incidence (Bonassi and Au 2002; HUMN: The International Collaborative Project on Micronucleus Frequency in Human Populations 2011).

Chromosome Aberrations Detection of chromosome aberrations is extensively used in molecular epidemiology as a biomarker of chromosomal damage, genome instability, and eventually of cancer risk. The detection of these altered chromosome formations from blood lymphocytes is used as part of the efforts in evaluating exposures to hazardous chemicals or suspected environmentally released and/or toxic compounds. Large prospective studies of cancer-free individuals have been conducted that demonstrate that chromosomal aberrations can predict cancer occurrence among the cohort 15 years later (Olden 1994; Bonassi et al. 2000; Hagmar et al. 2004).

Although very time consuming, the chromosomal aberration (CA) assay is considered one of the best validated biomarkers of early biological effects in population studies. This assay detects chromosomal abnormality at the metaphase stage of the cell cycle. Exposure-induced CAs have been found sometimes to be specific to the agent of exposure. This may allow for the CA assay to distinguish between radiation and chemical exposures. In a few cases, the assay has been able to specify certain chemical exposures. The distinctive CAs induced by various exposures have also permitted results of the assay to predict survival of cells with certain CAs (Bonassi and Au 2002). Fluorescence in situ hybridization and whole chromosome painting, a technique that uses chromosome-specific probes that are labeled with fluorescent dyes and hybridized to chromosomes to detect chromosome rearrangements and region-specific breaks, are two common techniques utilized in studies evaluating CAs (Bonassi and Au 2002).

Based on their review of findings from CA studies, Bonassi and Au (2002) provided four arguments indicating that CAs are predictive of the development of cancer: (i) chromosome rearrangements play an important role in the activation of proto-oncogenes and inactivation of tumor suppressor genes, (ii) congenital disease has been characterized with abnormally high CAs and increased malignancies, (iii) all neoplastic cells have been found to have alterations in the karyotype, some of which are highly specific for particular diagnostic categories, and (iv) carcinogenic chemicals tend to be clastogenic and clastogenicity is known to be associated with known human carcinogens. A review of biomarkers in molecular epidemiology studies with respect to CAs and their role in health risk assessment and prediction of cancer has been given by Bonassi and Au (2002).

Sister Chromatid Exchange During the S phase of the cell cycle, DNA is replicated, and each chromosome is present in a duplicated state with the two genetically identical chromatids joined together at the centromere. These two sister chromatids are readily apparent in late prophase or early metaphase of mitosis. Sister chromatid exchange (SCE) is the process wherein the two sister chromatids break and rejoin with one another, physically switching positions on the chromosome. Because the exchanges occur with exact precision to the DNA sequence, the sister chromatids stay genetically identical. No genetic information is altered during this process, and no mutations can be detected. It is important to note that these exchanges are natural events during cell replication. Usually, each cell can undergo three or four SCEs during each replication cycle. Even though this process is common, it can also be induced by mutagens and xenobiotics that form DNA adducts or that interfere with DNA replication processes and enzymes. The formation of SCEs has been correlated with the effectiveness of recombinational DNA repair and the induction of point mutations, gene amplification, and cytotoxicity.

The measurement and counting of SCEs has been applied in several in vitro genotoxic effects studies (Gu et al. 1981; Beckman and Nordenson 1986; Baldacci et al. 1997; Kasuba et al. 2000; Wagner et al. 2004; Bonina et al. 2008). An increased number of exchanges (15–100/cell) is considered to be an indication of increased cancer risk and genotoxic chemical exposure. Furthermore, in Bloom syndrome, an autosomal recessive disorder with defects in DNA repair by homologous recombination, pediatric patients have more than 160 SCEs per cell and have increased risk for developing cancers. This assay represents an integrated measure of chromosome-instability phenotypes and altered cellular viabilities caused by genetic defects and/or exogenous exposures to genotoxic agents.

Reporter Gene Mutations The HPRT (hypoxanthine phosphoribosyltransferase) gene mutation assay detects somatic mutations of the HPRT hemizygous gene in the X chromosome in post-thymic lymphocytes and is most commonly used for detection of mutation(s) in exposed populations (Bonassi and Au 2002). This cell cloning assay can distinguish between high mutant frequencies arising from clonal expansion of only a few mutated cells versus induction of a large number of mutated cells. Since the assay quantifies mutations in the HPRT gene which is only present on one allele, it is difficult to extrapolate mutant frequencies to genes present on both alleles. The HPRT assay allows for evaluation of the mutagenic potential of various compounds; however, the significance of these observations in terms of risk assessment for somatic mutation related diseases is difficult to evaluate (Bonassi and Au 2002). HPRT mutant frequencies have been utilized as a biomarker of effect in a study of polymer production workers, exposed to 1,3-Butadiene (Ammenheuser et al. 2001). HPRT mutant frequencies were found to be threefold greater in the high-exposure group. In addition, the assay has the utility to evaluate in utero exposures. This was evaluated in cord blood lymphocytes from healthy newborns in which the mutation frequency correlated with the presence of DNA adducts (Perera et al. 2002).

The glycophorin A mutation assay (GPA) detects damage in bone marrow cells. The mechanism of genotoxicity measured with this assay is the loss of alleles, which may explain why this assay is less sensitive than the HPRT assay (Bonassi and Au 2002). Fluorescent-labeled monoclonal antibodies and erythrocyte flow cytometry are used to detect the loss of one of the cell surface glycoproteins associated with the allele(s) of interest (Lee et al. 2002). A dose-dependent increase in the frequency of variants has been detected following exposure to genotoxic agents (Grant and Bigbee 1993). Further, Lee and colleagues (2002) found a significant association between GPA variant levels and urinary metabolite levels in a study of workers from an industrial waste incinerating plant, demonstrating the utility of the GPA as an early marker of biological effects of polycyclic aromatic hydrocarbon exposure.

Acquired Somatic Mutations in Genes There are several examples of studies considering somatic (acquired) gene mutations in relation to cancer. CDKN2A is a cell-cycle gene that is important in many cancers. Mutation of this gene, whether inherited or acquired, has been found in melanoma, lung cancer, among others. For example, p16^{INK4a} belongs to the family of cell-cycle-regulator enzymes called cyclin-dependent kinase inhibitors (CDKI), which bind to “cyclin-CDK” complexes and cause cell-cycle arrest in the G1 phase. p16^{INK4a} exerts its antiproliferative effects by binding to and inhibiting the actions of CDK 4 and 6 cyclin-dependent kinases.

Many studies in cancer have explored the importance of the integrity of this tumor suppressor protein and its activity. The INK4a gene is located at the chromosome locus 9p21, a region that has frequently been detected to undergo hemi- and homozygous deletion in human cancers. Both germline and somatic mutations of INK4a have been detected in familial melanoma and pancreatic adenocarcinoma. Somatic mutations of this gene were identified in Barrett’s esophagus-related adenocarcinomas, transitional cell carcinomas of the bladder, head and neck squamous cell carcinomas, and a variety of other malignancies (Caldas et al. 1994; Liu et al. 1995; Smith-Sørensen and Hovig 1996). Recent investigations have revealed another important molecular pathway of gene inactivation, hypermethylation of the INK4a-promoter region, which may be the predominant mechanism in some tumors (Belinsky et al. 2002). Tobacco smoke exposure can increase methylation levels in the INK4a-promoter region (Soria et al. 2002; Belinsky 2004).

Immunohistochemical staining (IHC) has been developed for p16^{INK4a} protein assessment which is applicable to fixed paraffin-embedded tumor tissues. Like all IHC assays, this technique has the advantage of detecting the p16^{INK4a} status irrespective of the mechanism of gene inactivation. For example, although there is a low frequency of INK4a mutations or allelic deletions in this tumor, nearly 20% of gastric carcinomas demonstrate reduction or loss of p16^{INK4a} expression by IHC, underscoring the importance of epigenetic and environmentally induced mechanisms in downregulating protein expression (Tamura 2006).

p53, another tumor suppressor gene, is expressed in every cell of the human body. It controls the synthesis of proteins that are involved in the regulation of cell division, DNA replication, cell differentiation, and in scheduling cell necrosis (Lutz and Nowakowska-Swirta 2002). p53 is a negative regulator of cell division and will block growth of cells and initiate repair mechanisms in cells with damaged DNA. Congenital defects and/or externally induced damage to this gene have extensive consequences on the maintenance of cellular homeostasis. This gene is of particular interest as it is has been found to be mutated in approximately half of human cancer cases (Lutz and Nowakowska-Swirta 2002).

Several studies investigating the presence of p53 gene mutations in lung cancer and normal tissue among smokers and non-smokers have found a correlation between the frequency of p53 mutations and the daily amount of smoking (reviewed in Vineis and Husgafvel-Pursianinen (2005)). Mutations in the p53 gene have been found to precede clinical symptoms by months to years in hepatic angiosarcoma patients exposed to vinyl chloride (Lutz and Nowakowska-Swirta 2002). In their review of p53, Lutz and Nowakowska-Swirta (2002) suggested that detecting mutations in this gene may serve as a way to identify early phases of carcinogenesis at the cellular level and thus become an early biomarker of health effect.

The finding that p53 mutations and changes in its expression form the basis of cancer processes has prompted molecular epidemiologists to use biomarkers in easily available cellular material or systemic liquids. Mutations in the suppressor gene p53 cause variations of cellular protein p53 concentration because the mutated protein accumulates in cancer tissues due to the loss of its DNA-binding capacity. Higher cellular protein p53 levels are associated with increased protein transfer to the extracellular liquid and to blood. It has been observed that increased blood serum protein p53 concentrations may have a prognostic value in early diagnosis of lung cancer (reviewed in Lutz and Nowakowska-Swirta (2002)). Elevated blood serum protein p53 concentrations have been observed also in patients with mesothelioma many years before the cancer became clinically overt (Hemminki et al. 1996).

46.5 Biomarkers of Susceptibility

Molecular epidemiology includes the identification of genetic and acquired susceptibility (such as DNA repair capacity – Berwick and Vineis 2000). Genetic epidemiology methods are well described in chapter ▸Statistical Methods in Genetic Epidemiology of this handbook. Association studies are the most common genetic epidemiology studies carried out these days. A large number of studies have been conducted on candidate genes on the basis of biochemical hypotheses regarding their involvement in carcinogen metabolism, DNA repair, or cell cycle. We will not address this issue here. Multiple meta-analyses and GWAS analyses (see below) are currently evaluating large numbers of SNPs (see Manolio and Collins (2009) for an overview of methods) and genetic loci that appear to be associated with disease. Although these studies tend to be large, including thousands of cases and thousands of controls, some smaller studies with very well-selected subjects have contributed to the understanding of genetic susceptibility (Klein et al. 2005). For example, in the study by Klein et al. (2005), a genome-wide screen of 96 cases and 50 controls identified an intronic and common variant in the complement factor H gene (CFH) that is clearly important in age-related macular degeneration (odds ratio $(OR) = 7.4$, 95% confidence interval (CI) = (2.9;19)).

46.5.1 Genome-Wide Association Studies (GWAS)

Genome-wide association studies (GWAS) are exploratory studies of genetic variation across the entire human genome designed to identify genetic associations with observable traits or the presence or absence of a certain disease of interest (Pearson and Manolio 2008). The basic premise is that the entire genome can be evaluated for variation and some SNPs will stand out as important risk factors for disease. Individuals with disease are compared to individuals without disease. As the genome is large and the number of SNPs is even larger, thousands of subjects are necessary to appropriately investigate associations. The agnostic approach of GWAS provides an advantage over candidate gene studies and has increased the exploratory potential of genetic analyses.

GWAS employ 500,000 to more than more than 2,500,000 SNP-genotyping platforms that provide cost-effective increases not only in the genetic information gained but also because they allow large populations to be studied. GWAS collect a large number of study participants with a disease or phenotypic trait of interest and a similarly large comparable, disease-free comparison group. The study sample originates usually from ongoing collaborative scientific efforts involving different institutions and even different continents. These studies take advantage of high-throughput genotyping technologies, automated sample collection, DNA isolation methods, and high-quality-control practices. They also employ statistical analyses to identify a relationship between certain SNPs that have passed assay-dependent detection thresholds and the presence of a disease or phenotypic trait. Tremendous laboratory and biostatistical effort has resulted in over hundreds of GWAS so far which no doubt will contribute to the knowledge base of molecular epidemiology around the world (McCarroll and Altshuler 2007). Accurate GWAS also try to replicate their findings in different population samples or in experimental animals, when the biological pathways allow mechanistic modeling or toxicological applications. These studies have also been used to identify SNPs associated with different gene expression patterns and certain phenotypic traits, like height and electrocardiographic QT intervals (Weedon et al. 2007, Arking et al. 2006).

Considering the “common disease, common variant” hypothesis, GWAS rely on SNPs as indicators of allelic variants that represent over 1–5% of each human genome. The National Cancer Institute has provided support for many cancer studies (breast, prostate, and colon) that have applied high-throughput powerful genomic analyses to reveal some of the key genetic factors that affect risk for these cancers. Through genetic characterization, followed by finer and finer mapping and analysis, scientists are able to identify common genetic aspects of the cancers being studied. Usually, genome-wide scanning is performed on an initial group of cases and controls, and then a smaller series of possible SNPs are evaluated as replicates in a second and maybe a third set of cases and controls. This multistaging study design was implemented to reduce the number of false-positive or false-negative findings (Skol et al. 2007). Furthermore, this methodology helps to reduce the genotyping cost of GWAS. Also, with carefully employed quality controls, the repeated genotyping provides the required validation, especially for unknown functional or intronic SNPs.

GWAS are prone to biases. One of the most important is population stratification. Population stratification is an important confounding effect where an underlying population structure contributes to false-positive associations with the disease being studied and occurs when the detected SNPs are also linked with unknown factors reflecting ethnicity or geographical origin of study participants. The large amount of data generated by GWAS is highly susceptible to false-positive associations. Effective statistical techniques have to be applied to reduce the occurrence of false-positives raised by the large number of multiple comparisons.

Another important limitation is that even though current technology used in GWAS captures the majority of common variants, they are determined based upon the concept of linkage disequilibrium or other statistical algorithms validated across the Human Genome International HapMap databases. These approaches build on our understanding that the human genome is organized by blocks of nucleotides called haplotypes. Haplotypes are inherited together; they form a linkage disequilibrium, and some SNPs within a given block define and explain or “tag” within block variability (Zhang et al. 2002). These tagging SNPs became popular in GWAS. Nevertheless, it is possible that some variants are still not captured by the applied genotyping chips although they represent potentially important but unknown susceptibility factors. Furthermore, it is imperative to consider that certain genetic variants might be important only in conjunction with exposures that initiate or modify expression of that gene (Ambrosone 2007). Without accounting for exposures to evaluate cancer or other chronic disease risks, we cannot successfully reveal the intricate patterns of gene-environment and even gene-gene interactions which account for a large proportion of cancer or chronic disease risk. “Next-generation GWAS” probably should incorporate more detailed analyses of common exposures (smoking, air pollutants, dietary patterns and supplements, xenobiotics (like common non-steroidal anti-inflammatory drugs (NSAIDs)), over-the-counter medications, alcohol, recreational drugs, etc.) influential in chronic disease etiology and pathogenesis.

46.5.2 Copy-Number Variation

Another interesting area of research that has emerged from the use of GWAS is the study of copy-number variations (CNVs) across the human and primate genomes (Dumas et al. 2007). Human populations have been found to have extensive polymorphisms not only in nucleotide sequences within genes but also in the form of additions or deletions resulting in variations in the number of copies of chromosomal segments (Hastings et al. 2009). A CNV is a segment of DNA in which the number of copies differs from that in a reference library (Pinto et al. 2007). Copy-number variants are as important as SNPs in determining individual differences between human beings (Kidd et al. 2008). The number of gene segments that differ from one individual to another varies for particular genetic regions due to deletion or duplication of DNA. The gene segments that are affected by the CNV processes can range from larger than 1 kilobase to up to many megabases. CNVs may either be inherited or caused by de novo mutation (Stankiewicz and Lupski 2010) and later many of them have become fixed in humans. Most of the CNVs are stretches of duplicated DNA that are more than 1 kilobase in size and share a sequence similarity that exceeds 90%. They are presented twice or more within the genome. It is thought that CNVs could have serious implications for the functionality of the gene and subsequently contribute to disease susceptibility (Gu and Lupski 2008).

Based on data generated from the Human Genome Project, approximately 12% of the human genome is subject to CNV, which is greater than the variation due to SNPs found to date (Sebat et al. 2004; Hastings et al. 2009). However, 75% of all CNVs detected have a prevalence of less than 3% in human populations. CNVs can arise both meiotically and somatically. Even identical twins can have different CNVs based on somatic events during their lives. More interestingly, tissues and organs from the same individual can vary in copy number, which indicates that probably CNVs are randomly incorporated in the genome. It is possible that genes with a higher frequency of CNVs are directly relevant to the immediate environment of the individual. CNVs are enriched in genes involved in olfaction, immunity, and secreted proteins. One example is the presence and protein amount of salivary amylase (*AMY1*) in diverse human populations (Perry et al. 2007) especially those with high starch consumption. It is believed that genes under recent selection pressure contain higher than average frequencies of non-synonymous mutations and CNVs. However, it is possible that because CNVs are randomly present, their functionality can change over time or in different environments, and their role within the human genome is continuously tested by evolutionary processes.

Copy-number variations require a change in chromosome structure. During cell division, two different single-stranded DNA structures get close to each other and join DNA sequences. Probable molecular events behind the mechanisms for de novo CNVs are fork stalling and template switching, both of which are considered frequent mistakes during genomic replication. Several methods are used to study how the structural change has arisen in humans; however, most of the molecular studies have used bacteria and yeast model organisms.

46.5.3 Mendelian Randomization

Establishing causal relationships between environmental exposures and common diseases is beset with problems of unresolved confounding, reverse causation, and selection bias that may result in spurious associations.

Mendelian randomization (Davey-Smith et al. 2004) is a method of using measured variation in genes of known function to examine the causal effect of a modifiable exposure on disease in non-experimental studies. An important focus of observational epidemiology is the identification of modifiable causes of common diseases that are of public health interest. In order to have firm evidence that a recommended public health intervention will have the desired beneficial effect, the observed association between the particular risk factor and disease must imply that the risk factor actually causes the disease. Well-known successes include the identified causal links between smoking and lung cancer, and between blood pressure and stroke. However, there have also been notable failures when identified exposures were later shown by randomized controlled trials (RCTs) to be non-causal. For instance, it has now been shown that hormone replacement therapy (HRT) will not prevent cardiovascular disease, as was previously thought. HRT may even have other adverse health effects (Rossouw et al. 2002). The reason for such spurious findings in observational epidemiology is most likely to be confounding by social, behavioral, or physiological factors which are difficult to control for and particularly difficult to measure accurately. Moreover, many findings cannot be replicated by RCTs for ethical reasons.

Mendelian randomization is a method that allows one to test for, or in certain cases to estimate, a causal effect from observational data in the presence of confounding factors by using common genetic polymorphisms with well-understood effects on exposure patterns (e.g., propensity to drink alcohol; Davey-Smith et al. 2004) or effects that mimic those produced by modifiable exposures (e.g., raised blood cholesterol; Katan 1986). Importantly, the genotype must only affect the disease status indirectly via its effect on the exposure of interest. Because genotypes are assigned randomly when passed from parents to offspring during meiosis, if we assume that choice of mate is not associated with genotype (panmixia), then the population genotype distribution should be unrelated to the confounders that typically plague observational epidemiology studies. In this regard, Mendelian randomization can be thought of as a "natural" RCT. From a statistical perspective, it is an application of the technique of instrumental variables, with genotype acting as an instrument/proxy for the exposure of interest. As with all genetic epidemiology studies, there are problems associated with the need for large sample sizes, the non-replication of findings, and the lack of relevant functional information on genetic variants (refer to chapter ▸Statistical Methods in Genetic Epidemiology of this handbook for more information on genetic epidemiology). In addition to these problems, genetic findings may be confounded by other genetic variants in linkage disequilibrium with the variant under study or by population stratification. Furthermore, pleiotropy of effect of a genetic variant may result in null associations, as may canalization of genetic effects. If correctly conducted and carefully interpreted, Mendelian randomization studies can provide useful evidence to support or reject causal hypotheses linking environmental exposures to common diseases.

46.6 Methodological Issues in Using Biomarkers

It is widely accepted that environmental exposure measurements in cancer epidemiology require well-developed and validated exposure assessment methodologies (Wild 2009; Vineis and Perera 2007; Gyorffy et al. 2008). Before newly developed high-throughput technologies can be used on a regular basis in exposure or risk assessment, they will require extensive validation.

Rundle (2006) described common methodological limitations in using biomarkers, including the use of target versus surrogate tissues, modifying effects of the disease on DNA adduct levels, the use of inappropriate statistical analyses, and small sample sizes in published data. These problems make it difficult to assess whether lack of an association between the presence of carcinogen-induced DNA adducts and cancer reflects a lack of association between the xenobiotic exposure of interest and cancer or mirrors problems in methodology (Rundle 2006). A focus on measurement of adduct levels in tissues collected years prior to a cancer diagnosis (i.e., in benign biopsy specimens) may resolve some of these uncertainties.

The measurement of DNA-based markers depends on the type and amount of DNA available: genomic DNA (including mitochondrial DNA) or tumor DNA. Amount of DNA available and methods of collection and storage all greatly affect the quality of the assays and thus the appropriate application. Microarray technology was made available to researchers in the late 1980s. The main principle of this technology is based on the Southern blot assay in which DNA fragments containing a specific DNA sequence of interest that serve as a probe to hybridize to a complementary DNA (cDNA), which has been amplified using fluorescent primers. If the cDNA contains the complementary sequence, it will hybridize with the probe, and the fluorescence is measured. The quality of the DNA becomes an issue when considering the area of interest within the DNA sequence as well as the requirements for amplification. This becomes particularly important when designing studies using stored DNA samples in which the quality of the DNA is often not known.

46.7 Risk Assessment

The main goal in molecular epidemiological research is to explore and explain relationships between environmental factors (including diet and behavioral exposures) and health outcomes. The extensive scientific inquiry of molecular epidemiology uses and appreciates the far-reaching framework of risk assessment used in regulatory risk management and policy making (World Health Organization (WHO) 2012; US Environmental Protection Agency (EPA) 1989).

The first phase of risk assessment is *exposure assessment* (for a full description, see chapter ▸Exposure Assessment of this handbook and the specific section in chapter ▸Environmental Epidemiology for general principles of scientific environmental risk assessment in this handbook) based on results of pertinent toxicological and epidemiological studies. This step also involves a description of the chemical or biological agent's behavior and interactions within the human body. Applicable molecular markers developed for this purpose may also be considered in molecular epidemiological studies linking organ-, tissue-, and cellular-level changes to disease development. Animal models are frequently used in this phase of scientific inquiry especially for evaluations of potential toxic exposures. Research then links laboratory and field observations of adverse health effects with exposure to particular agents (e.g., leukemia as a result of benzene exposure).

The second step of risk assessment is to evaluate whether dose-effect and sometimes more accurately dose-response relationships exist by describing and quantifying the relationship between exposure or absorbed dose and related health outcomes. Often this component of the assessment is hindered by the lack of appropriate methods necessary to extrapolate data from high-dose (e.g., occupational exposure settings) to low-dose exposure (e.g., general air pollution exposure conditions) or from animal models to humans (Yassi et al. 2001; Friis 2007).

High-risk groups are important to identify in epidemiology, particularly in relation to environmental health hazards. Fetuses, infants, young children, elderly people, pregnant women, the nutritionally deprived, and people with existing chronic conditions or genetic disorders are often more sensitive to physical, chemical, and biological hazards. Since individuals have varied exposures and susceptibility, the standards for protective measures should be set based on the most vulnerable population groups, that is, public health policies should attempt to provide protective measures to those who are most susceptible.

Risk characterization is important for policy development, bringing together exposure assessment and potential dose-effect associations for estimating the incidence and severity of the potential adverse health effects. It involves exposure characterization, dose estimation, and prediction of lifetime risk, including susceptibility of subgroups. These steps lead to risk management, that is, the development and implementation of regulatory actions aimed to reduce risk and apply effective protective measures for the population at large.

46.8 Biorepositories

As a vital component of biomarker-related molecular epidemiological studies, biorepositories have been established in order to provide tissue and blood samples that are kept together with demographic and clinical data, including information regarding clinical diagnosis and response(s) to therapeutic options. Biorepositories involve collections of blood or tissue for genetic research that can be linked to medical, genealogical, or lifestyle information about a specific population, where the material is gathered using a specific consent process for future undetermined genetic and molecular research (Maschke 2005). Biorepositories lend themselves to the ultimate goal of molecular epidemiological studies which is to identify indicators of disease susceptibility or prognostic markers to reduce disease incidence and mortality. Thus, biorepositories are an invaluable resource for studies of cancer etiology, progression, and development of new biomarkers of precursors of disease (Ambrosone 2006).

Population-based biorepositories consist of specimens from large populations of individuals. Attempts to create such large repositories usually require a coordinated effort at the national level (Maschke 2005; Goodman et al. 2005). This type of repository for tissue, blood compartments, urine, DNA, or other biological material is sometimes initiated for surveillance or as part of genome studies and has the advantage of providing an unbiased sampling frame that is useful for examining genes and protein targets as therapeutic and/or diagnostic tools (Goodman et al. 2005). In the case of cancer research, population-based tissue repositories are ideal for testing new molecular classification schemes for cancer, for validation of new biological markers of malignancy, for prognosis of disease progression, for assessment of therapeutic targets, and for measurement of allele frequencies of cancer-associated genetic polymorphisms in representative samples (Goodman et al. 2005).

The US National Cancer Institute (NCI) created a national program of population-based cancer registries in 1973, the Surveillance, Epidemiology and End Results (SEER) program (National Cancer Institute 2011). SEER collects information on cancer incidence, survival, and prevalence from 26 specific geographical areas within the United States (Seer 2013). Recently, this cancer registry has expanded to include a collection of formalin-fixed, paraffin-embedded tissues from cancer patients at the population level. This repository is unique in the sense that the large-scale collection is performed at the population level and yet these tissues constitute specimens from many different types of one disease, cancer (Goodman et al. 2005). Further, Goodman and colleagues suggest that this type of tissue resource may provide a foundation for molecular epidemiological studies of cancer in the USA.

Another example of a large-scale population-based tissue resource in the USA is the Cooperative Human Tissue Network (CHTN) which was initiated by the Cancer Diagnosis Program of NCI in 1987 as a means of providing basic and applied scientists from academia and industry with access to human cancer tissues. This biorepository was established with the intention of accelerating the advancement of discoveries in cancer diagnosis and treatment. Currently, six institutions throughout the USA are funded to procure and distribute tissues to biomedical researchers throughout the USA and Canada. Materials are obtained from surgical resections, autopsies, and remnant serum/plasma and other biofluids (urine and buffy coat). The CHTN continuously updates a list of requested tissues from regional investigators, and when a tissue request matches a surgical resection, the excess specimen not needed for diagnosis is prepared and preserved according to the specified needs of the investigator (National Cancer Institute 2012). CHTN provides investigators with tissues that can be age-, sex-, and race-matched based on their needs; this allows for robust data collection for molecular epidemiological studies (case-control matching is discussed further in chapter ▶Case-Control Studies of this handbook). Even though this resource is valuable for molecular epidemiology studies, samples will most likely be used to generate hypotheses because classical epidemiological information, for example, exposure data, cannot be obtained through the tissue repositories.

The International Agency for Research on Cancer (IARC 2012) hosts the database and biorepository for the largest study of diet and health ever undertaken, the European Prospective Investigation into Cancer and Nutrition (EPIC project). The EPIC project has recruited over 520,000 participants in ten European countries and was designed to investigate the relationships between diet, nutritional status, lifestyle, and environmental factors and the incidence of cancer and other chronic diseases. Scientists included in the EPIC project are based in 23 centers throughout the ten participating European countries and have utilized the available individual data (dietary, lifestyle), follow-up data (cancer, mortality, and other endpoint data), and biospecimens including plasma, serum, red blood cells, and buffy coats, to conduct investigations that have resulted in over 768 publications up to 2008. The study is jointly coordinated at IARC and Imperial College in London.

The UK Biobank (2012) is another large-scale European effort designed to demonstrate the effect of lifestyle, environment, and genes on the health of 500,000 people aged 40–69 from the United Kingdom. This study is designed to collect blood, urine, standard measurements, and lifestyle information from participants that can be utilized by approved researchers over the next 20–30 years. The overall aim of this study is to help researchers develop new and better ways to prevent, diagnose, and treat chronic illnesses (UK Biobank 2012).

Biorepositories consisting of tissues from subgroups within a population can be for example created as part of collaborations between public clinics/health-care facilities and biotech companies. These studies are defined as convenience studies and can be subject to selection bias (Goodman et al. 2005). Potential problems exist among biorepositories: the type and timing of tissues collected and stored. Tissue is either collected during surgery, treatment, or in the course of a diagnostic procedure or alternatively collected only for the purpose of research (van Veen 2006). Residual tissue can be used in research provided consent is obtained, or the tissues are not linked to any personal identifiers. Use of tissues collected explicitly for research is dependent upon whether appropriate consent was obtained.

Although biorepositories are an excellent resource for molecular epidemiological studies, access to samples and inconsistencies in the quality of the samples or problems with sample integrity can hinder utilization of these resources (Ambrosone 2006). The long-term value of biospecimens is dependent upon proper collection and handling of specimens, accurate collection of appropriate data about the study participant, and specimen quality (Vaught 2006). Despite appropriate collection, processing, and storage of specimens, variations prior to analysis could influence the validity of the results.

The International Society for Biological and Environmental Repositories (ISBER), which was established in 1999, has developed best practice guidelines for biorepositories. The inception of these guidelines has initiated a focused attention on the important aspects of proper biospecimen handling as well as legal and ethical standards and regulations that govern specimen collection (Vaught 2006). The ISBER Biospecimen Working Group has compiled a list of publications that focus on examination and validation of various experimental designs in accordance with biospecimen research (International Society for Biological and Environmental Repositories 2012).

Ethical Issues Collection, analysis, and storage of biospecimens and medical information from individuals entail several legal and ethical requirements. Ideas regarding the regulation of tissue repositories and the existence of laws to enforce this control vary widely at the international level. In particular, issues relating to informed consent, access to research results, the use of archived samples, and the ownership of biological samples as well as intellectual property of discoveries and inventions have been debated. Ideally, an international regulatory framework would be the most appropriate entity to address these issues; however, until its establishment, laws, regulations, and ethics advisory board guidelines at the national level govern the collection, storage, and research use of biological samples (Maschke 2005).

Informed consent provides individuals with information related to the research and the ability to decide whether they are willing to have their samples collected, stored, and studied. Voluntary informed consent is a core ethical principle of human research ethics (Maschke 2005).

Archived biospecimens are ideal for studies evaluating long-term survival; biomarkers of interest can be evaluated to determine whether their presence or absence predicts survival from a specific disease years after the specimen was collected. An area of recent debate is whether the use of archived specimens that were not previously consented for future research is ethically sound (Maschke 2005).

Regarding ownership of biological specimens and intellectual property rights to discoveries and inventions resulting from research with biospecimens, the general consensus in the international and national law has been that there are no property rights on the human body (Maschke 2005). For additional examples of ethical issues arising in epidemiological studies, refer to chapter ►Ethical Aspects of Epidemiological Research of this handbook.

46.9 Conclusions

Conventional epidemiology, based on classical research methods and tools such as interviews and questionnaires, has achieved extremely important goals and contributed tremendously to public health policy and regulatory changes. Even a difficult issue such as the relationship between air pollution and chronic disease has been successfully dealt with by time-series analysis and other methods not based in the laboratory. The added value of molecular techniques coupled with solid and carefully executed epidemiological design can be most appreciated in complex and chronic disease research.

In addition to its positive features, the limitations of molecular epidemiology should also be acknowledged: the complexity and cost of many laboratory methods, with partially unknown levels of measurement error or interlaboratory variability; the scanty knowledge of the sources of bias and confounding; in some circumstances, the lower degree of accuracy (e.g., urinary cotinine compared to questionnaires of smoking history); and the uncertain biological meaning of some markers, as in the case of some types of adducts or some early response markers. The timing of the exposure and the biological measurement often do not reflect the cumulative nature of the exposure which is usually dependent upon the biological pathway (e.g., vitamin D serum levels or arsenic in toenails).

Molecular epidemiology is not distinct from traditional epidemiology, but represents a development that aims at achieving specific scientific goals: (a) a better characterization of exposures, particularly when levels of exposure are very low or different sources of exposure are integrated into a single measure; (b) the study of gene-environment interactions; and (c) the use of markers of early response, in order to overcome the main limitations of chronic disease epidemiology, that is, the relatively low frequency of specific forms of disease and the long latency period between exposure and the onset of disease.

References

Albertini RJ, Anderson D, Douglas GR, Hagmar L, Hemminki K, Merlo F, Natarajan AT, Norppa H, Shuker DEG, Tice R, Waters MD, Aitio A (2000) IPCS guidelines for the monitoring of genotoxic effects of carcinogens in humans. Mutat Res 463:111–172

Ambrosone CB (2006) Sample collection, processing, and storage for large-scale studies: biorepositories to support cancer research. Cancer Epidemiol Biomarkers Prev 15(9):1574

Ambrosone CB (2007) The promise and limitations of genome-wide association studies to elucidate the causes of breast cancer. Breast Cancer Res 9:114

Ammenheuser MM, Bechtold WE, Abdel-Rahman SZ, Rosenblatt JI, Hastings-Smith DA, Ward JB (2001) Assessment of 1,3-Butadiene exposure in polymer production workers using HPRT mutations in lymphocytes as a biomarker. Environ Health Perspect 109:1249–1255

Angerer J, Ewers U, Wilhelm M (2007) Human biomonitoring: state of the art. Int J Hyg Environ Health 210:201–228

Ankathil R, Jyothish B, Madhavan J, Nair MK (1999) Deficient DNA repair capacity: a predisposing factor and high risk predictive marker in familial colorectal cancer. J Exp Clin Cancer Res 18(1):33–37

Arking DE, Pfeufer A, Post W, Kao WH, Newton-Cheh C, Ikeda M, West K, Kashuk C, Akyol M, Perz S, Jalilzadeh S, Illig T, Gieger C, Guo CY, Larson MG, Wichmann HE, Marbán E, O'Donnell CJ, Hirschhorn JN, Kääb S, Spooner PM, Meitinger T, Chakravarti A (2006) A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization (QT interval). Nat Genet 38(6):644–651

Baldacci S, Carrozzi L, Viegi G, Giuntini C (1997) Assessment of respiratory effects of air pollution: study design on general population samples. J Enviorn Pathol Toxicol Oncol 16(2–3):77–83

Beckman L, Nordenson I (1986) Interaction between some common genotoxic agents. Hum Hered 36(6):397–401

Belinsky SA (2004) Gene-promoter hypermethylation as a biomarker in lung cancer. Nat Rev Cancer 4:707–717

Belinsky SA, Snow SS, Nikula KJ, Finch GL, Tellez CS, Palmisano WA (2002) Aberrant CpG island methylation of the p16(INK4A) and estrogen receptor genes in rat lung tumors induced by particulate carcinogens. Carcinogenesis 23:335–339

Berwick M, Vineis P (2000) Markers of DNA repair and susceptibility to cancer in humans: an epidemiologic review. J Natl Cancer Inst 92:874–897

Berwick M, Song Y, Jordan R, Brady MS, Orlow I (2001) Mutagen sensitivity as an indicator of soft tissue sarcoma risk. Environ Mol Mutagen 38:223–226

Betsou F, Barnes R, Burke T, Coppola D, DeSouza Y, Eliason J, Glazer B, Horsfall D, Kleeberger C, Lehmann S, Prasad A, Skubitz A, Somiari S, Gunter E (2009) Human biospecimen research: experimental protocol and quality control tools. Cancer Epidemiol Biomarkers Prev 18(4):1017–1025

Bonassi S, Au WW (2002) Biomarkers in molecular epidemiology studies for health risk prediction. Mutat Res 511:73–86

Bonassi S, Hagmar L, Stromberg U, Montagud AH, Tinnerberg H, Forni A, Heikkila P, Wanders S, Wilhardt P, Hansteen IL, Knudsen LE, Norppa H (2000) Chromosomal aberrations in lymphocytes predict human cancer independently of exposure to carcinogens. Cancer Res 60:1619–1625

Bonassi S, Fenech M, Lando C, Lin YP, Ceppi M, Chang WP, Holland N, Kirsch-Volders M, Zeiger E, Ban S, Barale R, Bigatti MP, Bolognesi C, Jia C, Di Giorgio M, Ferguson LR, Fucic A, Lima OG, Hrelia P, Krishnaja AP, Lee TK, Migliore L, Mikhalevich L, Mirkova E, Mosesso P, Müller WU, Odagiri Y, Scarffi MR, Szabova E, Vorobtsova I, Vral A, Zijno A (2001) HUman MicroNucleus project: international database comparison for results with the cytokinesis-block micronucleus assay in human lymphocytes: I. Effect of laboratory protocol, scoring criteria, and host factors on the frequency of micronuclei. Environ Mol Mutagen 37(1):31–45

Bondy ML, Kyritsis AP, Gu J, de Andrade M, Cunningham J, Levin VA, Bruner JM, Wei Q (1996) Mutagen sensitivity and risk of gliomas: a case-control analysis. Cancer Res 56(7):1484–1486

Bondy ML, Wang L-E, El-Zein R, de Andrade M, Selvan MS, Bruner JM, Levin VA, Yung WKA, Adatto P, Wei Q (2001) Gamma-radiation sensitivity and risk of glioma. J Natl Cancer Inst 93:1553–1557

Bonina FP, Puglia C, Frasca G, Cimino F, Trombetta D, Trigali G, Roccazzello A, Insiriello E, Rapisarda P, Saija A (2008) Protective effects of a standardized red orange extract on air pollution-induced oxidative damage in traffic police officers. Nat Prod Res 22(17):1544–1551

Caldas C, Hahn SA, da Costa LT, Redston MS, Schutte M, Seymour AB, Weinstein CL, Hruban RH, Yeo CJ, Kern SE (1994) Frequent somatic mutations and homozygous deletions of the p16 (MTS1) gene in pancreatic adenocarcinoma. Nat Genet 8(1):27–32

Caporaso N (1999) Genetics of smoking-related cancer and mutagen sensitivity. J Natl Cancer Inst 91(13):1097–1098

Cloos J, Spitz MR, Schantz SP, Hsu TC, Zhang ZF, Tobi H, Braakhuis BJ, Snow GB (1996) Genetic susceptibility to head and neck squamous cell carcinoma. J Natl Cancer Inst 88(8):530–535

Cloos J, Nieuwenhuis EJ, Boomsma DI, Kuik DJ, van der Sterre ML, Arwert F, Snow GB, Braakhuis BJ (1999) Inherited susceptibility to bleomycin-induced chromatid breaks in cultured peripherial blood lymphocytes. J Natl Cancer Inst 91(13):1097–1098

Cloos J, Leemans CR, van der Sterre ML, Kuik DJ, Snow GB, Braakhuis BJ (2000) Mutagen sensitivity as a biomarker for second primary tumors after head and neck squamous cell carcinoma. Cancer Epidemiol Biomarkers Prev 9(7):713–717

Cloos J, de Boer WP, Snel MH, van den Ijseel P, Ylstra B, Leemans CR, Brakenhoff RH, Braakhuis BJ (2006) Microarray analysis of bleomycin-exposed lymphoblastoid cells for identifying cancer susceptibility genes. Mol Cancer Res 4(2):71–77

Collins AR, Cadet J, Moller L, Poulsen HE, Vina J (2004) Are we sure we know how to measure 8-oxo-7,8-dihydroguanine in DNA from human cells? Arch Biochem Biophys 423:57–65

Colquhoun DR, Goldman LR, Cole RN, Gucek M, Mansharamani M, Witter FR, Apelberg BJ, Halden RU (2009) Global screening of human cord blood proteomes for biomarkers of toxic exposure and effect. Environ Health Perspect 117:832–838

Davey-Smith G, Harbord R, Ebrahim S (2004) Fibrinogen, C-reactive protein and coronary heart disease: does Mendelian randomization suggest the associations are non-causal? QJM 97(3):163–166

Davis B (2005) Growing pains for metabolomics. Scientist 19(8):25–28

Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR, Sikela JM (2007) Gene copy number variation spanning 60 million years of human and primate evolution. Genome Res 9:1266–1277

El-Zein R, Bondy ML, Wang LE, de Andrade M, Sigurdson AJ, Bruner JM, Kyritsis AP, Levin VA, Wei Q (2001) Risk assessment for developing gliomas: a comparison of two cytogenetic approaches. Mutat Res 490(1):35–44

Erdei E, Lee SJ, Wei Q, Wang LE, Song YS, Bovbjerg D, Berwick M (2006) Reliability of mutagen sensitivity assay: an inter-laboratory comparison. Mutagenesis 21:261–264

ESCODD (European Standards Committee on Oxidative DNA Damage) (2002) Comparative analysis of baseline 8-oxo-7,8-dihydroguanine in mammalian cell DNA, by different methods in different laboratories: an approach to consensus. Carcinogenesis 23:2129–2133

Friis RH (2007) Essentials of environmental health. Jones and Bartlett, Sudbury

Fry RC, Navasumrit P, Valiathan C, Svensson JP, Hogan BJ, Luo M, Bhattacharya S, Kandjanapa K, Soontararuks S, Nookabkaew S, Mahidol C, Ruchirawat M, Samson LD (2007) Activation of inflammation/NF-kappa B signaling in infants born to arsenic-exposed mothers. PLoS Gent 3:2180–2189

Goodman MT, Hernandez BY, Hewitt S, Lynch CF, Coté TR, Frierson HF Jr, Moskaluk CA, Killeen JL, Cozen W, Key CR, Clegg L, Reichman M, Hankey BF, Edwards B (2005) Tissues from population-based cancer registries: a novel approach to increasing research potential. Hum Pathol 36:812–820

Grant SG, Bigbee WL (1993) In vivo somatic mutation and segregation at the human glycophorin A (GPA) locus: phenotypic variation encompassing both gene-specific and chromosomal mechanisms. Mutat Res 288:163–172

Gu W, Lupski JR (2008) CNV and nervous system diseases – what's new? Cytogenet Genome Res 123(1–4):54–64

Gu ZW, Sele B, Jalbert P, Vincent M, Vincent F, Marka C, Chmara D, Faure J (1981) Induction of sister chromatid exchange by trichloroethylene and its metabolites. Toxicoil Eur Res 3(2): 63–67

Gyorffy E, Anna L, Kovacs K, Rudnai P, Schoket B (2008) Correlation between biomarkers of human exposure to genotoxins with focus on carcinogen-DNA adducts. Mutagenesis 23:1–18

Hagmar L, Stromberg U, Bonassi S, Hansteen IL, Knudsen LE, Lindholm C, Norppa H (2004) Impact of types of lymphocytechromosomal aberrations on human cancer risk: results from Nordic and Italian cohorts. Cancer Res 64:2258–2263

Hastings PJ, Lupski JR, Rosenberg SM, Ira G (2009) Mechanisms of change in gene copy number. Nat Rev Genet 10:551–564

Hemminki K, Partanen R, Koskinen H, Smith S, Carney W, Brandt-Rauf PW (1996) The molecular epidemiology of oncoproteins: serum p53 protein in asbestosis patients. Chest 109:22S-26S

Hsu TC, Johnston DA, Cherry LM, Ramkissoon D, Schantz SP, Jessup JM, Winn RJ, Shirley L, Furlong C (1989) Sensitivity to genotoxic effects of bleomycin in humans: possible relationship to environmental carcinogenesis. Int J Cancer 43:403–409

HUMN: The International Collaborative Project on Micronucleus Frequency in Human Populations (2011) http://ehs.sph.berkeley.edu/holland/humn/. Accessed on 11 Feb 2012

IARC (International Agency for Research on Cancer) European Prospective Investigation into Cancer and Nutrition – EPIC project (2012) http://epic.iarc.fr/centers/iarc.php. Accessed on 26 Jan 2012

International Programme on Chemical Safety (1999) Environmental health criteria 210. World Health Organization, Geneva. http://www.inchem.org/documents/ehc/ehc/ehc210.htm. Accessed on 4 Feb 2012

International Society for Biological and Environmental Repositories (ISBER) (2012) http://www.isber.org/wg/. Accessed on 11 Feb 2012

Kang H, Chen IM, Wilson CS, Bedrick EJ, Harvey RC, Atlas SR, Devidas M, Mullighan CG, Wang X, Murphy M, Ar K, Wharton W, Borowitz MJ, Bowman WP, Bhojwani D, Carroll WL, Camitta BM, Reaman GH, Smith MA, Downing JR, Hunger SP, Willman CL (2010) Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric B-precursor acute lymphoblastic leukemia. Blood 115(7):1394–1405

Kapp AV, Jeffrey SS, Langerød A, Børresen-Dale AL, Han W, Noh DY, Bukholm IR, Nicolau M, Brown PO, Tibshirani R (2006) Discovery and validation of breast cancer subtypes. BMC Genomics 7:231

Kasuba V, Rozgaj R, Sentija K (2000) Cytogenetic changes in subjects occupationally exposed to benzene. Chemosphere 40(3):307–310

Katan MB (1986) Apolipoprotein E isoforms, serum cholesterol, and cancer. Lancet 1(8479): 507–508

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453(7191):56–64

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 3(5720):385–389

Lee KH, Lee J, Ha M, Choi JW, Cho SH, Hwang ES, Park CG, Strickland PT, Hirvonen A, Kang D (2002) Influence of polymorphism of GSTM1 gene on association between glycophorin A mutant frequency and urinary PAH metabolites in incineration workers. J Toxicol Environ Health A 65:355–363

Liu L, Lassam NJ, Slingerland JM, Bailey D, Cole D, Jenkins R, Hogg D (1995) Germline p16INK4A mutation and protein dysfunction in a family with inherited melanoma. Oncogene 11(2):405–412

Lutz W, Nowakowska-Swirta E (2002) Gene p53 mutations, protein p53, and anti-p53 antibodies as biomarkers of cancer process. Int J Occup Med Environ Health 15:209–218

Manolio TA, Collins FS (2009) The HapMap and genome-wide association studies in diagnosis and therapy. Annu Rev Med 60:443–456

Maschke KJ (2005) Navigating an ethical patchwork-human gene banks. Nat Biotechnol 23(5):539–545

McCarroll SA, Altshuler DM (2007) Copy-number variation and association studies of human disease. Nat Genet 39(7)(suppl):S37–S42

National Cancer Institute (2011) Surveillance Epidemiology and End Results (SEER). http://seer.cancer.gov/. Accessed on 4 Feb 2012

National Cancer Institute (2012) Cooperative Human Tissue Network (CHTN)-Human Biospecimens for Research. http://www.chtn.org/. Accessed on 11 Feb 2012

Neri M, Fucic A, Knudsen LE, Lando C, Merlo F, Bonassi S (2003) Micronuclei frequency in children exposed to environmental mutagens: a review. Mutat Res 544:243–254

Olden K (1994) Mutagen sensitivity as a biomarker of genetic predisposition to carcinogenesis. J Natl Cancer Inst 86(22):1660–1661

Olive PL, Banath JP (2006) The comet assay: a method to measure DNA damage in individual cells. Nat Protoc 1:23–29

Pearson TA, Manolio T (2008) How to interpret a genome-wide association study. JAMA 299(11):1335–1344

Perera F, Hemminki K, Jedrychowski W, Whyatt R, Campbell U, Hsu Y, Santella R, Albertini R, O'Neill JP (2002) In utero DNA damage from environmental pollution is associated with somatic gene mutation in newborns. Cancer Epidemiol Biomarkers Prev 11:1134–1137

Perera FP, Tang D, Rauh V, Tu YH, Tsai WY, Becker M, Stein JL, King J, Del Priore G, Lederman SA (2007) Relationship between polycyclic aromatic hydrocarbon-DNA adducts, environmental tobacco smoke, and child development in the world trade center cohort. Environ Health Perspect 115:1497–1502

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, Carter NP, Lee C, Stone AC (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet 39(10):1256–1260

Pinto D, Marshall C, Feuk L, Scherer SW (2007) Copy-number variation in control population cohorts. Hum Mol Genet 16(2):R168–R173

Psychogios N, Hau DD, Peng J, Guo AC, Mandal R, Bouatra S, Sinelnikov I, Krishnamurthy R, Eisner R, Gautam B, Young N, Xia J, Knox C, Dong E, Huang P, Hollander Z, Pedersen TL, Smith SR, Bamforth F, Greiner R, McManus B, Newman JW, Goodfriend T, Wishart DS (2011) The human serum metabolome. PLoS One 6(2):e16957

Robertson KD, Wolffe AP (2000) DNA methylation in health and disease. Nat Rev Genet 1:11–19

Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, Jackson RD, Beresford SA, Howard BV, Johnson KC, Kotchen JM, Ockene J, Writing Group for the Women's Health Initiative Investigators (2002) Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. JAMA 288(3):321–333

Rundle A (2006) Carcinogen-DNA adducts as a biomarker for cancer risk. Mutat Res 600(1–2):23–26

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305(5683):525–528

Seer (2013) http://seer.cancer.gov/

Sigurdson AJ, Bonday ML, Hess KR, Toms SA, Kyristsis AP, Gu J, Wang LE, Wang X, Adatto P, Bruner JL, Yung WK, Levin VA, Wei Q (1998) Gamma-ray mutagen sensitivity and survival in patients with glioma. Clin Cancer Res 4(12):3031–3035

Skol AD, Scott LJ, Abecasis GR, Boehnke M (2007) Optimal designs for two-stage genome-wide association studies. Genet Epidemiol 31(7):776–788

Smith-Sørensen B, Hovig E (1996) CDKN2A (p16INK4A) somatic and germline mutations. Hum Mutat 7(4):294–303. Review.

Soria JC, Rodriguez M, Liu DD, Lee JJ, Hong WK, Mao L (2002) Aberrant promoter methylation of multiple genes in bronchial brush samples from former cigarette smokers. Cancer Res 62:351–355

Spira A, Beane J, Shah V, Liu G, Schembri F, Yang X, Palma J, Brody JS (2004) Effects of cigarette smoke on the human airway epithelium cell transcriptome. Proc Natl Acad Sci 101: 10143–10148

Spitz MR, Hoque A, Trizna Z, Schantz SP, Amos CI, King TM, Bondy ML, Hong WK, Hsu TC (1994) Mutagen sensitivity as a risk factor for second malignant tumors following malignancies of the upper aerodigestive tract. J Natl Cancer Inst 86(22):1681–1684

Stankiewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. Ann Rev Med 61:437–455

Tamura G (2006) Alterations of tumor suppressor and tumor-related genes in the development and progression of gastric cancer. World J Gastroenterol 12(2):192–198

UK Biobank (2012) http://www.ukbiobank.ac.uk/. Accessed on 4 Feb 2012

US Environmental Protection Agency (EPA) (1989) Risk assessment guidance for superfund Volume I, human health evaluation manual. Office of Emergency and Remedial Response, Washington DC. EPA/540/1-89/002

van Veen BE (2006) Human tissue bank regulations. Nat Biotechnol 24(5):496–497

Vaught JB (2006) Biorepository and biospecimen science: a new focus for CEBP. Cancer Epidemiol Biomarkers Prev 15(9):1572–1573

Vineis P, Gallo V (eds) (2007) Epidemiological concepts of validation of biomarkers for the identification/quantification of environmental carcinogenic exposures. ECNIS, environmental cancer risk, nutrition and individual susceptibility, Lodz, Poland. http://www.ecnis.org/index.php?option=com_content&task=view&id=626&Itemid=135

Vineis P, Husgafvel-Pursianinen K (2005) Air pollution and cancer: biomarker studies in human populations. Carcinogenesis 26:1846–1855

Vineis P, Perera F (2007) Molecular epidemiology and biomarkers of etiologic cancer research: the new in light of the old. Cancer Epidemiol Biomarkers Prev 16:1954–1965

Wagner KH, Jurss A, Zamrembach B, Elmadfa I (2004) Impact of antiseptics on radical metabolism, antioxidant status and genotoxic stress in blood cells: povidone-iodine versus octenidine dihydrochloride. Toxicol Vitro 18(4):411–418

Weedon MN, Lettre G, Freathy RM, Lindgren CM, Voight BF, Perry JR, Elliott KS, Hackett R, Guiducci C, Shields B, Zeggini E, Lango H, Lyssenko V, Timpson NJ, Burtt NP, Rayner NW, Saxena R, Ardlie K, Tobias JH, Ness AR, Ring SM, Palmer CN, Morris AD, Peltonen L, Salomaa V; Diabetes Genetics Initiative; Wellcome Trust Case Control Consortium, Davey Smith G, Groop LC, Hattersley AT, McCarthy MI, Hirschhorn JN, Frayling TM (2007) A common variant of HMGA2 is associated with adult and childhood height in the general population. Nat Gen 39(10):1245–1250

Wild CP (2009) Environmental exposure measurement in cancer epidemiology. Mutagenesis 24:117–125

Wild CP, Vineis P, Garte S (eds) (2008) Molecular epidemiology of chronic diseases. Wiley, West Sussex

World Health Organization (WHO) (2012) Risk assessment. http://www.who.int/topics/risk_assessment/en/. Accessed on 4 Feb 2012

Wu X, Lippman SM, Lee JJ, Zhu Y, Wei QV, Thomas M, Hong WK, Spitz MR (2002) Chromosome instability in lymphocytes: a potential indicator of predisposition to oral premalignant lesions. Cancer Res 62(10):2813–2818

Yassi A, Kjellstrom T, de Kok T, Guidotti T (2001) Basic environmental health. World Health Organization. Oxford University Press, Oxford

Zhang K, Calabrese P, Nordborg M, Sun F (2002) Haplotype block structure and its applications to association studies: power and study designs. Am J Hum Genet 71:1386–1394

Zheng YL, Loffredo CA, Yu Z, Jones RT, Krasna MJ, Alberg AJ, Yung R, Perlmutter D, Enewold L, Harris CC, Shields PG (2003) Bleomycin-induced chromosome breaks as a risk marker for lung cancer: a case-control study with population and hospital controls. Carcinogenesis 24(2):269–274

Zhu Y, Spitz MR, Hsu TC, Wu X (2002) Genetic instability of specific chromosomes associated with a family history of cancer. Cancer Genet Cytogenet 136(1):73–77

Clinical Epidemiology and Evidence-Based Health Care

47

Holger J. Schünemann and Gordon H. Guyatt

Contents

H.J. Schünemann (✉) • G.H. Guyatt
Faculty of Health Sciences, Departments of Clinical Epidemiology and Biostatistics and of Medicine, McMaster University, Hamilton, ON, Canada

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_30,

47.1 Introduction

This chapter will begin with providing a brief overview of the history of clinical epidemiology and describe its relation with evidence-based medicine. Clinical epidemiology differs from classical epidemiology in that clinical epidemiology supports other basic medical sciences such as biochemistry, anatomy, and physiology because it facilitates their application in research through formulation of sound clinical research methods and, thus, puts these disciplines into clinical context. Therefore, clinical epidemiology goes beyond clinical trials. We will describe this concept in the following paragraphs (see Sects. 47.1.1–47.1.3). The following sections include case scenarios that facilitate the introduction of the key concepts about developing clinical questions, using diagnostic tests, evaluating therapy, appraising systematic reviews, developing guidelines, and making clinical decisions. By describing clinical epidemiology in this way, the relation to evidence-based health care (which in the authors' view encompasses all fields of medicine and associated clinical sciences) will become clear.

47.1.1 Brief History of Clinical Epidemiology

Sackett provides an astute historical summary of the development of clinical epidemiology in his tribute in memory of Alvan Feinstein (Sackett 2002). Sackett's account gives John Paul credit for introducing the term clinical epidemiology describing it as the "new basic science for preventive medicine" (Paul 1938; Sackett 2002). Feinstein and Sackett made major contributions to the field of clinical epidemiology. Sackett founded, in 1966, the first clinical epidemiology research unit at the University at Buffalo, New York, USA, and in 1967 moved to Hamilton, Canada, to establish a department of clinical epidemiology and biostatistics at McMaster University, which he served as chair from 1968 to 1973. This institution, in part through the International Clinical Epidemiology Network (INCLEN), has trained numerous clinical epidemiologists, some of whom have taken on chair positions themselves. It has grown to host one of the largest graduate programs worldwide in health research methodology with over 180 Master's and PhD students. Today there are departments and units of clinical epidemiology throughout the world, though the development in some jurisdictions has been slower than in others. For example, it was in this millennium that the first department of clinical epidemiology was founded in the German-speaking countries of Europe (Basel, Switzerland, H. Bucher, personal communication), although professorships of clinical epidemiology existed in these countries for some time.

47.1.2 A Definition of Clinical Epidemiology

Although seminal, Paul's simple description of clinical epidemiology was perhaps not sufficient in helping investigators and clinicians understand the principles underlying the term clinical epidemiology. Articles and textbooks have provided further definitions. Feinstein portrayed clinical epidemiology as investigating "the occurrence rates and geographical distribution of disease; the pattern of natural and post-therapeutic events that constitute varying clinical courses in the diverse spectrum of disease; and the clinical appraisal of therapy. The contemplation and investigation of these or allied topics constitute a medical domain that can be called clinical epidemiology" (Feinstein 1968; Sackett et al. 1991). Sackett defined clinical epidemiology as "the application, by a physician who provides direct patient care, of epidemiological and biostatistical methods to the study of diagnostic and therapeutic processes in order to effect an improvement of health" (Sackett and Winkelstein 1967; Sackett 1969, 2002). Fletcher et al. (1996) described clinical epidemiology as the science of making predictions about individual patients by counting clinical events in similar patients, using strong scientific methods for studies of groups of patients to ensure that the predictions are accurate. Weiss (1996) defined clinical epidemiology as the study of variation in the outcome of illness and of the reasons for that variation. Despite the numerous definitions, one might argue that by providing the subheading "A Basic Science for Clinical Medicine" to their textbook "Clinical Epidemiology," Sackett and colleagues provided a pithy definition that not only turned the wheel back to John Paul but widened it to all areas of clinical medicine by replacing *preventive medicine* with *clinical medicine* (Sackett et al. 2000).

Definitions are inevitably limited, and in-depth understanding requires a more comprehensive discussion. We characterize clinical epidemiology by focusing on its purpose: to ensure that clinicians' practice and decision-making is evidence-based. Clinical decision-making requires answering questions about diagnosis, therapy, prevention, and harm, providing estimates of prognosis and obtaining unbiased and precise estimates of intervention effects. Clinical epidemiology supports other basic medical and health sciences such as biochemistry, anatomy, and physiology because it facilitates their application in research through formulation of sound clinical research methods and, thus, puts these disciplines into the health care context. Thus, clinical epidemiology provides the integrative force of medical science and medical practice.

47.1.3 Clinical Epidemiology and Evidence-Based Medicine

When working optimally, clinical epidemiologists communicate results of investigations in ways that clinicians can readily apply in practice. Clinical epidemiology provides the evidence for management decisions resulting in more good than harm. In fact, the methodological principles of clinical epidemiology question the millennia old credo of "primum non nocere." Clinical decision-making demands that the commonplace interpretation as "never do harm" is impossible to adhere to. Every decision in life, including the simplest ones, but particularly health care decisions, comes with downsides. The smallest perceptible downside is the time to consider the decision, but there are also always opportunity costs that result from giving up an alternative action. Clinicians' and patients' action, therefore, require the explicit consideration of both desirable and undesirable consequences and assigning due considerations depending on the magnitude and importance of the consequences. Understanding of health research methodology and its application in clinical epidemiology is critical for these considerations and for weighing benefits and downsides of decisions. One of us, therefore, described the guiding principle for health care decision-making as to ensure that there is, in summary, more benefit than harm; in other words, the credo should be "to do no net harm" ("primum non net nocere") (Schunemann 2011).

Clinicians should use best evidence to consider these consequences for clinical decision-making. Thus, clinical epidemiology and evidence-based practice are closely linked. Clinical epidemiology grounds health care research in the mission to deliver optimal care to individual patients. As it turns out, clinicians optimally applying the best evidence to their patient care must understand the basic concepts of clinical epidemiology (Bhandari et al. 2003). At the same time, while clinical epidemiology grounds the clinical investigators' viewpoint, evidence-based medicine (EBM) provides the framework for application of research findings in clinical practice, including a framework for clinical decision-making. In the next sections of this discussion, we will describe the basics of clinical epidemiology methods and how insights from clinical practice may enlighten clinical epidemiologists in their work from "study to bedside."

47.1.4 The History and Philosophy of Evidence-Based Medicine

When one of us first coined the term evidence-based medicine (EBM) in an informal residency training program document, he described it as "an attitude of enlightened skepticism toward the application of diagnostic, therapeutic, and prognostic technologies in their day-to-day management of patients" (Guyatt 1991, 2002a, b). Through a series of articles published by the evidence-based medicine working group, the term as well as the philosophy of EBM became well known (Evidence-Based-Medicine-Working-Group 1992; Oxman et al. 1993). A Medline search revealed 7 citations including the term "evidence based medicine" in 1993, 3,904 citations in 2002 (when we worked on the first edition for this book), and 7,573 in 2011 (when we worked on the current edition of this book). The history of evidence-based health care has been summarized in a book (Daly 2005). While the Department of Clinical Epidemiology and Biostatistics at McMaster University is widely regarded the home of EBM, EBM became quickly established across the world. For example, Sackett spent several years in the UK to develop the Oxford Centre for Evidence-based Medicine.

EBM evolved out of the efforts of clinicians with research methodology training – that is, clinical epidemiologists – to apply their particular insights and approaches to solving clinical problems. In contrast to the traditional paradigm of clinical practice, EBM acknowledges that intuition, unsystematic clinical experience, and pathophysiologic rationale are important but not sufficient for

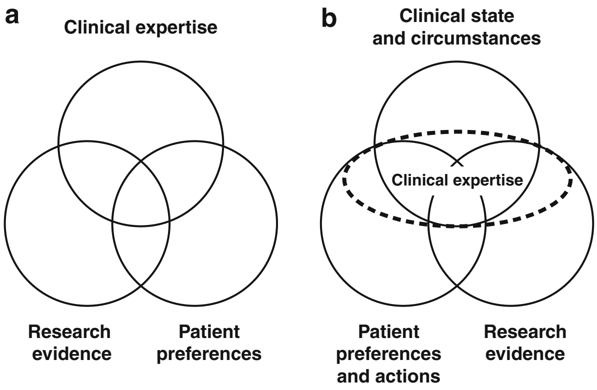


Fig. 47.1 (**a**) and (**b**) Former models of evidence-based decision-making (Reproduced with permission from Haynes et al. (2002a, b))

making the best clinical decisions. Although it acknowledges the importance of clinical experience, EBM postulates that optimal clinical decision-making requires the integration of this experience with the circumstances (setting), patients' values and preferences, and estimates of effects of interventions or diagnostic strategies. In 2002, Haynes and colleagues described a model for making evidence-based clinical decisions (Haynes et al. 2002a, b). That model focuses on evidence as research evidence and requires an integration of other factors to make the best decisions. Since then, the field of practice guideline development has led us to develop an extended model of evidence-based decision-making.

In that model, there was increasing recognition (Fig. 47.1a, b) that clinical expertise is an overarching pillar of supporting best decision-making but that these decisions should be based on the best research evidence. The better recognition of the importance of patients' state and circumstances as well as values and preferences supplemented the original model. However, the emphasis was on research evidence about effects and a clear delineation of what constitutes evidence to support decision-making was missing.

EBM places a lower value on authority than the traditional medical paradigm and explicitly includes patients' and society's values in the clinical decision-making process. Patients or their proxies must always trade the benefits, harm, and costs associated with alternative treatment strategies and in doing so must consider values and preferences.

To achieve the integration of research results in clinical practice, EBM proposes a formal set of rules to help clinicians interpret and apply evidence. Clinical epidemiologists have, by and large, developed these rules, but they have been refined in the two decades of evidence-based medicine. These rules are characterized by a hierarchy of evidence: that is, the confidence in research results is greatest if systematic error (bias) is lowest and decreases if bias is more likely to play a role. However, they have moved from simple, now outdated

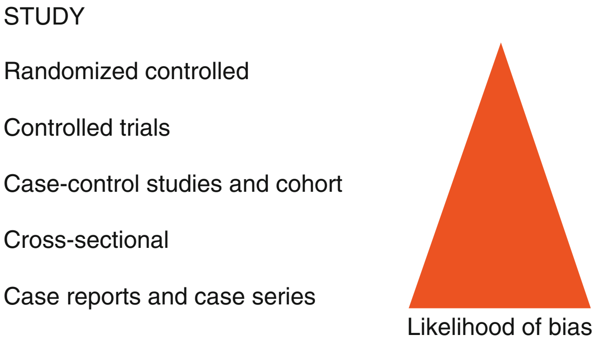


Fig. 47.2 Depicts the hierarchy of quality of evidence. As the research design becomes more rigorous (moving from *bottom* to *top*), the likelihood of bias of individual studies decreases

hierarchies as shown in the figure in the prior version of this chapter (Fig. 47.2) to balanced approaches that include considerations about "the evidence about the evidence."

Although randomized and controlled study designs provide the highest quality of evidence from a perspective of single studies, EBM is not a science of randomized controlled trials. Rather, because higher-quality evidence often is not available, EBM acknowledges that a large body of highly relevant evidence comes from observational studies. That is, to answer clinical questions clinicians will often depend on observational studies in their evidence-based practice. Therefore, the practice and application of EBM requires an understanding and critical evaluation of all study designs. Clinical epidemiology provides the necessary toolbox for this evaluation. For example, clinical epidemiologists of the Cochrane Collaboration, an international organization dedicated to making up-to-date and accurate information about the effects of health care readily available worldwide, have provided important insights into the conduct of systematic reviews and meta-analysis that inform clinicians and patients choices. It produces and disseminates systematic reviews of health care interventions and promotes the search for evidence. It can be accessed at www.cochrane.org. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group has developed an approach to assessing the quality of a body of evidence for a given question, ideally based on highly credible systematic reviews (Atkins et al. 2004; Guyatt et al. 2008b, 2011a, b; Schünemann et al. 2011c). We will describe this approach in more detail below.

In a further evolved model for evidence-based decision-making, increasing emphasis is placed on the recognition that evidence about the various factors that influence decision-making can come from various sources and that a single patient in the clinical decision-making context provides evidence about her own values which requires integration with other pieces of evidence. Without doubt must the highest demands made toward obtaining any evidence with methods that are least vulnerable to bias. This new model also includes considerations about the best implementation and knowledge integration methods as a critical part of medical decision-making (Fig. 47.3).

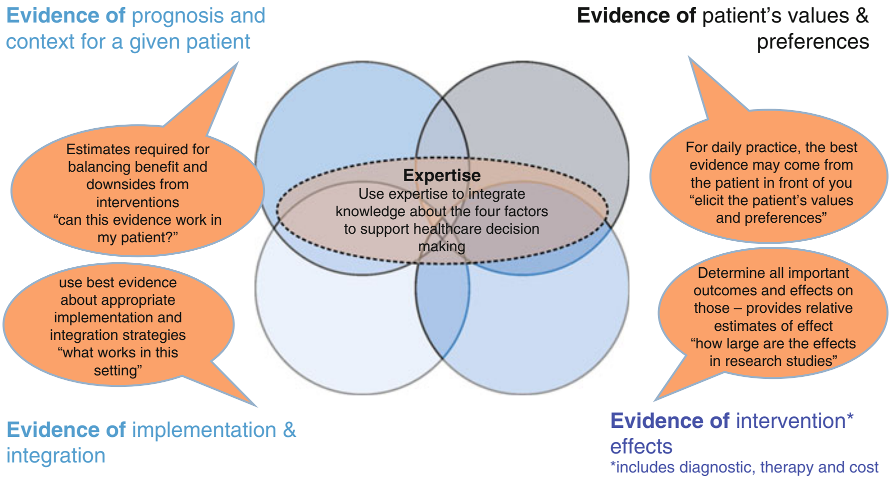

Evidence of prognosis and context for a given patient
Evidence of patient's values & preferences
Estimates required for balancing benefit and downsides from interventions "can this evidence work in my patient?"
Expertise
Use expertise to integrate knowledge about the four factors to support healthcare decision making
For daily practice, the best evidence may come from the patient in front of you "elicit the patient's values and preferences"
use best evidence about appropriate implementation and integration strategies "what works in this setting"
Determine all important outcomes and effects on those – provides relative estimates of effect "how large are the effects in research studies"
Evidence of implementation & integration
Evidence of intervention* effects
*includes diagnostic, therapy and cost

Fig. 47.3 Evolved model of evidence-based decision-making

47.2 Case Scenario

Example 47.1.
Imagine you are the attending physician on ward rounds with your team. The senior resident presents the case of a 64-year-old woman who came to the emergency room one early morning with left-sided chest pain lasting for 15 minutes. The pain was severe enough to awaken her. She also had to sit up in her bed because of difficulties with getting her breath. Finally, her symptoms became so severe that she called an ambulance.

You immediately think that this woman has had an acute myocardial infarction. However, other diagnoses such as pulmonary embolism, pneumonia, pericarditis, asthma, and a severe case of gastroesophageal reflux disease also come to mind.

The resident continues with her presentation and tells you that the pain radiated to her left arm and neck and you further hear that she had another episode of similar pain in the ambulance. You feel that this information confirms your early intuition and that it makes a diagnosis of myocardial infarction more likely.

An electrocardiogram (EKG) in the emergency room was unremarkable and cardiac enzymes drawn at arrival to the emergency room were borderline elevated (troponin I, a marker for myocardial injury, was 1.5 microgram per milliliter (μg/ml)). Her chest x-ray was normal.

You now think that the diagnosis might be one of acute coronary syndrome, perhaps unstable angina or a myocardial infarction without EKG changes, and you continue to entertain pulmonary embolism, pneumonia, and pericarditis as alternative – although less likely – diagnosis. The key decision you face is whether to admit the patient to hospital, possibly to a cardiac care unit, or to send her home with provision for subsequent investigation, perhaps an exercise test.

The patient's past medical history includes diabetes mellitus type 2. Her lipid profile is within the limits set by the last available update of the National Cholesterol Education

Program Guidelines (Grundy et al. 2004), and there is no significant family history of cardiovascular disease. She takes an oral hypoglycemic agent and 325 mg of aspirin daily as recommended by her physician. She has had similar chest pain over the past year when vacuuming her home. However, she did not mention these complaints to her physician because the discomfort always resolved after a few minutes of rest. The patient has no history of cough, wheezing, indigestion, heartburn, or changes in her bowel habits. Your own physical examination of the patient following the resident's presentation shows an anxious patient, but there are no abnormal findings on physical examination.

Your team concludes that the probability that the presentation represents acute coronary syndrome is at least 50%. While you are discussing the patient and her further management, a second set of laboratory results shows that the troponin I is elevated at 4.1 μg/l.

At this point, you feel that a myocardial infarction without ST-segment elevation on the EKG (a non-ST-segment elevation myocardial infarction – NSTEMI) is the most likely diagnosis and together with your team you consider further management. This discussion focuses on which antiplatelet agents should be used in this patient.

47.3 Formulating a Clinical Question

Using research evidence to guide clinical practice requires formulating sensible clinical questions (Oxman et al. 1993; Richardson et al. 1995; McKibbon et al. 2002). For most questions, the key components are the patients, the intervention or exposure, comparison interventions (or exposure), and the outcomes (Table 47.1).

The clinical scenario of an older diabetic women presenting with chest pain potentially generates several clinical questions (about her diagnosis, appropriate therapy, prevention of future events, prognosis). We will use some of these questions to demonstrate how clinical epidemiology helps solve clinical problems and can make practice evidence-based.

Table 47.1 Formulating the clinical question

Component	Explanation
Population or patients	Who are the relevant patients?
Interventions or exposures	What are the management strategies clinicians are interested in? For example, diagnostic test, drugs, toxins, nutrients, and surgical procedures.
Comparison (or control) intervention or exposures	What is the comparison, control, or alternative intervention clinicians are interested in? For questions about therapy or harm, there will always be a comparison or control (including doing nothing, placebo, alternative active treatment, or routine care). For questions about diagnosis, there may be a comparison diagnostic strategy (for example, troponin I compared to creatine kinase MB in the diagnosis of myocardial infarction).
Outcome	What are the patient-important outcomes of the therapy, diagnostic test, or exposure clinicians are interested in?

47.4 Diagnosis

Based on the framework for developing a clinical question, we will start with a question about diagnosis:

Population:	In women with chest pain typical for angina pectoris
Intervention/exposure:	What is the test performance of troponin I serum levels?
Outcome:	To predict myocardial infarction and associated adverse outcomes (congestive heart failure, death, serious arrhythmia, or severe ischemic pain) in the next 72 hours

The process of diagnosis is a complex cognitive task. There are different approaches to making a diagnosis, but pattern recognition, which is also known as the gestalt method, and logical reasoning play an important role (Sox et al. 1988; Sackett et al. 1991; Glass 1996). Clinicians always look for clues that help them establish a diagnosis, although with increasing clinical experience this process becomes increasingly subconscious. Some clues make a diagnosis more likely; other clues or the absence of certain clues makes a diagnosis less likely (Ladenheim et al. 1987). In the scenario described at the beginning of this chapter, the first clue was the presence of left-sided chest pain awaking the patient at night. This clue suggested that the patient might suffer from a cardiac problem. The presence of shortness of breath strengthened this suspicion but brought other possible diagnosis into consideration (asthma, pneumonia, and pulmonary embolus).

When clinicians use clues offered by clinical history, symptoms, signs, or test results, they routinely, if often subconsciously, apply probabilities associated with these clues. For example, the presence of chest pain makes a heart attack more likely than no chest pain. Thus, a first step in making a diagnosis is to assign probabilities to the contemplated diagnoses. Clinicians then group the findings into coherent clusters, such as left-sided (location) chest (heart or lungs) pain (symptom). These clusters inform the differential diagnoses. The differential diagnoses in the case scenario included acute myocardial infarction, pulmonary embolism, pneumonia, asthma, or gastroesophageal reflux disease. In the next step of making a diagnosis, the clinician incorporates new information, which lowers or increases the relative likelihood of the differential diagnoses. The process is therefore sequential. The presence of pain radiating to the left arm increased the probability of coronary heart disease and the absence of cough and gastrointestinal symptoms lowered the likelihood of pneumonia and gastrointestinal disease.

47.4.1 Establish the Framework for Bayesian Thinking for Diagnosis

As described above, the process of diagnosis can take place on a subconscious level in which the clinician estimates the associated probabilities and relative likelihoods or it can employ explicit probabilities and relative likelihoods. For some clinical problems, such as diagnosing pulmonary embolism, a clinician's intuition is good. The Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED) study has shown this (PIOPED-Investigators 1990). Even when intuition is reasonably accurate, use of exact numbers generated by empirical studies can improve clinical decisions (Diamond and Forrester 1979; Diamond et al. 1980, 1981; Dolan et al. 1986). The latter approach of using explicit information is based on epidemiological and biostatistical concepts, but both the intuitive and the explicit approaches are founded in Bayesian theory, because both approaches depend on probabilities that are altered by subsequent information (Ledley and Lusted 1959; Bernardo and Adrian 1994; Berry 1996; Diamond 1999). Using the Bayesian approach in the diagnostic process, the clinician starts with a certain probability (often called the pretest probability) of a disease being present. Then, based on clues from the history, physical exam, or test results, the clinician modifies this probability into another probability (often called the posttest probability).

47.4.2 Choosing the Right Test

The best test would be one that excludes or confirms a diagnosis beyond doubt. Using an ideal test, no patient would have the disease if the test is negative and all patients with a positive test would have the disease. For our example, if troponin I was the perfect test, one could assume that a troponin I level $\geq 2\ \mu g/l$ proves beyond doubt that the patient has an acute myocardial infarction or will suffer a serious clinical event associated with acute coronary syndrome in the next 72 h, and a level $<2\ \mu g/l$ establishes that the patient does not have an acute myocardial infarction and will not suffer a serious event. Unfortunately, most information that clinicians obtain in clinical practice comes with uncertainty, and tests that definitively distinguish between disease and no disease are few and far between.

The typical cut-off value for troponin I in clinical practice is $2.0\ \mu g/l$ (Meier et al. 2002). However, astute clinicians would not dismiss a diagnosis of myocardial infarction in our scenario after the first troponin I level was $<2.0\ \mu g/l$, because both EKG and biomarkers may be what is typically defined as normal even when disease is present. Furthermore, they are aware that the troponin I may be normal early and may rise subsequently. What clinicians expect of a good test is that results change the probability sufficient to confirm or exclude a diagnosis. If a test result moves the probability below a threshold at which the disease is very unlikely and downsides associated with the treatment outweigh any anticipated benefit, then no further testing and no treatment are indicated. We call this probability the test threshold. If, on the other hand, the test result moves the probability of disease above a threshold at which one would not further test because disease is highly probable and one would start treatment, we have found the treatment threshold. This scenario shows how a clinician can estimate the probability of disease and then compare disease probability to these two thresholds (Fig. 47.4).

In a clinical context in which the pretest probability of a particular diagnosis is above the treatment threshold, further confirmatory testing that raises the probability further would not be helpful. On the other end of the scale, for a disease with a pretest probability below the test threshold, further exclusionary testing lowering

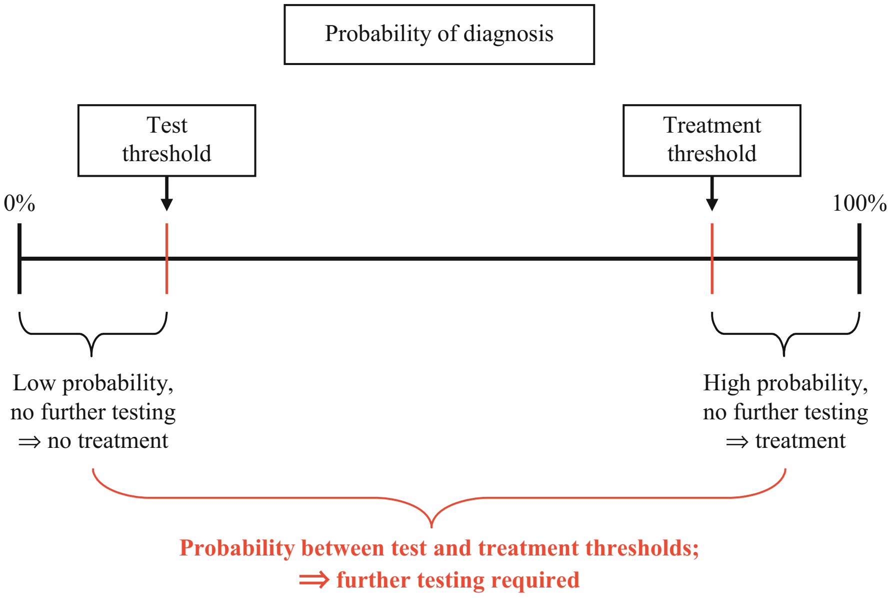

Probability of diagnosis
Test threshold
Treatment threshold
0%
100%
Low probability, no further testing ⇒ no treatment
High probability, no further testing ⇒ treatment
Probability between test and treatment thresholds; ⇒ further testing required

Fig. 47.4 Test and treatment thresholds

the probability would not be useful. When the probability is between the test and treatment thresholds, testing will be diagnostically useful. Test results are of greatest value when they shift the probability across either threshold.

What determines our treatment thresholds? If adverse effects of treatment are frequent and severe, clinicians choose a higher treatment threshold. For example, because a diagnosis of pulmonary embolism involves long-term anticoagulation with appreciable bleeding risk, clinicians are very concerned about falsely labeling patients. The invasiveness of the next test will also impact on the threshold. If the next test (such as a computed tomography angiography) does not have important side effects and generates few false-positives, clinicians are ready to choose a high treatment threshold. Clinicians are more reluctant to institute an invasive test associated with risks to the patient, such as a pulmonary angiogram, and this will drive their treatment threshold downward. That is, clinicians are more inclined to accept a risk of a false-positive diagnosis because a higher treatment threshold necessitates putting more patients through the risky test.

Accordingly, the more serious a missed diagnosis, the lower we will set our test threshold. Since a missed diagnosis of a pulmonary embolus could be fatal, clinicians are inclined to set their diagnostic threshold low. At the same time, the risks associated with the next test we are considering have an influence on where to set the test threshold. If the risks are low, clinicians will be comfortable with a very low diagnostic threshold. The higher the risks, the more the threshold rises.

47.4.3 Likelihood Ratios

The center of any diagnostic process is a patient presenting with a constellation of symptoms and signs. Consider two patients with chest pain and shortness of breath in whom the clinician suspects a myocardial infarction without findings suggestive of pneumonia, airflow obstruction, pulmonary embolism or heart failure, or other conditions. One patient is the 64-year-old woman described in the clinical scenario and the other is a 24-year-old man with a history of anxiety disorder. Clinicians would agree that the probabilities of myocardial infarction for these two patients – that is, their pretest probabilities – are very different. In the woman described in the scenario, the probability is high; in the young man, it is low. Consequently, even if both patients had borderline elevated troponin I levels of 1.0 μg/l at presentation, management is likely to differ between the two. An informed clinician might well treat the elderly woman immediately with aspirin, clopidogrel, anticoagulation, and percutaneous coronary interventions but order further investigations in the young man.

One can draw two conclusions from these considerations. First, regardless of the results of the troponin I test, they do not definitively establish whether myocardial infarction is in fact the underlying disease or whether the patient will suffer a serious event associated with an acute coronary syndrome. What they do accomplish is to alter the pretest probability of the condition, yielding a new posttest probability. The direction and magnitude of this change from pretest to posttest probability are determined by the test's properties. The test property of greatest value is the likelihood ratio.

Hill and colleagues (2004) investigated the diagnostic properties of troponin I as an early marker of acute myocardial infarction or acute coronary syndrome with serious sequelae in the next 72 h in patients who did not have definitively diagnostic EKG changes. The investigators found 20 individuals with a serious cardiac outcome by the reference standard and 332 individuals who did not (Table 47.2). For all patients, troponin I tests were classified into four levels: <0.5 μg/l, 0.5 to <2.0 μg/l, 2.0 to <10.0 μg/l, and ≥10.0 μg/l. Several questions arise.

How likely is a substantially elevated (≥10.0 μg/l) troponin I among people who suffered adverse outcomes? Table 47.2 illustrates that 3 of 20 (or approximately 15%) people with adverse outcomes had troponin I levels ≥10.0 μg/l. How often is the same test result, a positive troponin I, found among patients in whom high-risk acute coronary syndrome was suspected but ruled out? The answer is 4 out of 332 (or approximately 1.2%). The ratio of these two likelihoods is the likelihood ratio (LR); for a highly elevated troponin I test, it equals 0.15/0.012 (or 12.5). In other words, a highly elevated troponin I is 12.5 times as likely to occur in a patient with – in contrast to without – an ultimate adverse outcome.

In a similar fashion, one can calculate the likelihood ratios for troponin I values of <0.5 μg/l, 0.5 to <2.0 μg/l, and 2.0 to <10.0 μg/l. This calculation involves answering two questions: First, how likely it is to obtain a given test result (e.g., a troponin I < 0.5 μg/l) among people with the target disorder (myocardial infarction)? Second, how likely it is to obtain the same test result (again, a troponin

Table 47.2 Test properties of early troponin I testing in myocardial infarction or ischemia-associated adverse outcomes

	Myocardial Infarction or other adverse outcomes				
	Present | proportion		**Absent | proportion**		**Likelihood ratio (95% CI)**
Test results					
≥ 10.0 μg/l	3	3/20 = 0.15	4	4/332 = 0.012	12.5 (3.0, 51.9)
2.0 – <10.0 μg/l	2	2/20 = 0.10	5	5/332 = 0.015	6.6 (1.4, 32.1)
0.5 – < 2.0 μg/l	3	3/20 = 0.15	20	20/332 = 0.06	2.5 (0.8, 7.7)
<0.5 μg/l	12	12/20 = 0.60	303	303/332 = 0.910	0.7 (0.5, 0.9)
Total	20		332		

I $< 0.5\,\mu g/l$) among people without the target disorder? For this troponin I test result, the likelihoods are 12/20 (0.60) and 303/332 (0.91), respectively, and their ratio (the likelihood ratio) is 0.7.

Thus, the likelihood ratios indicate by how much a given diagnostic test result will raise or lower the pretest probability of the target disorder. A likelihood ratio of 1 indicates that the posttest probability is identical to the pretest probability. Likelihood ratios above 1.0 increase the probability that the target disorder is present, and the higher the likelihood ratio, the greater is this increase. Likelihood ratios below 1.0 decrease the probability of the target disorder, and the smaller the likelihood ratio, the greater is the decrease in probability and the smaller is its final value.

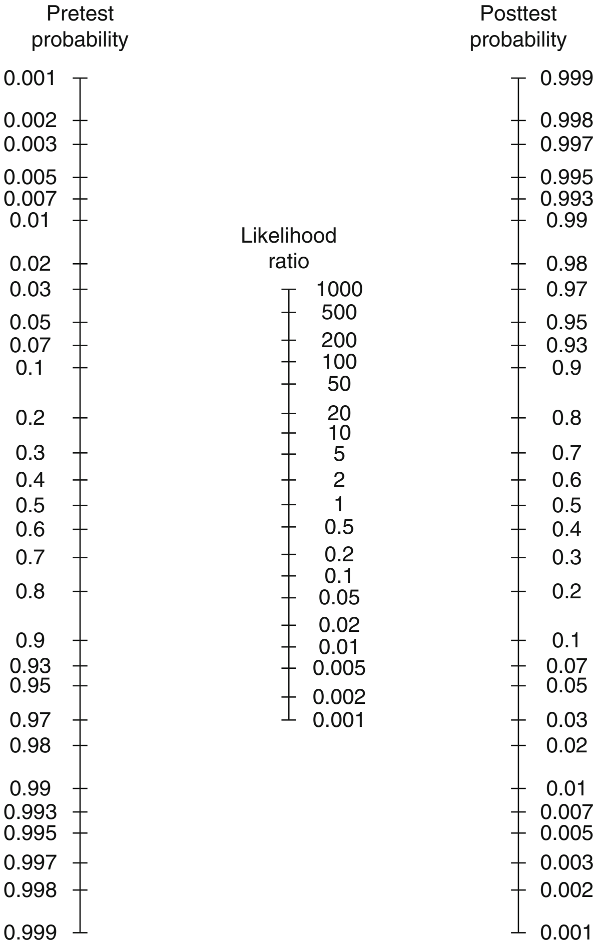
Users of likelihood ratios often ask “What are good likelihood ratios for a test?” The answer is that day-to-day clinical practice lets clinicians gain understanding and their own sense of interpretation and there is a strong dependence on the pretest probability, but one can consider the following as a rough guide:

1. Likelihood ratios of >10 or <0.1 generate large and often conclusive changes from pre- to posttest probability.
2. Likelihood ratios of 5–10 and 0.1–0.2 generate moderate shifts in pre- to posttest probability.
3. Likelihood ratios of 2–5 and 0.5–0.2 generate small (but sometimes important) changes in probability.
4. Likelihood ratios of 1–2 and 0.5–1 alter probability to a small (and rarely important) degree.

How can clinicians use likelihood ratios to move from pretest to posttest probability? Unfortunately, one cannot combine likelihoods directly, such as one can combine probabilities or percentages. Their formal use requires converting pretest probability to odds, multiplying the result by the likelihood ratio, and then converting the posttest odds into a posttest probability. Although this calculation is relatively straightforward for an experienced user, it can be time consuming and, fortunately, there is an easier way.

Figure 47.5 shows a nomogram proposed by Fagan that performs the conversions and allows simple transition from pre- to posttest probability (Fagan 1975). The line

Fig. 47.5 Likelihood ratio (Fagan) nomogram (copyright 1975 Massachusetts Medical Society. All rights reserved) (Reproduced with permission from the Massachusetts Medical Society)



on the left-hand side represents the pretest probability, the middle line represents the likelihood ratio (LR), and the line on the right-hand side depicts the resulting posttest probability. One can obtain the posttest probability by anchoring a straight line at the pretest probability and rotating it until it lines up with the likelihood ratio of the relevant test result.

If we assumed a pretest probability of 50% (see below) for the elderly woman with multiple risk factors and we applied the LR associated with a troponin I of 1.0 μg/l to the nomogram (connecting 0.5 or 50% on the left with an LR of approximately 2.5 on the middle line and extending it through the right line), we would obtain a posttest probability of approximately 70% (or 0.7). If we assumed a pretest probability of 1 in 1,000 or 0.1% for the young man and applied the same LR, the posttest probability remains very low at between 0.2% and 0.3%.

To further explain the application of LRs, let us assume the two patients had troponin levels of 0.3 μg/l. Applying the associated LR (0.7) to the pretest probability of 50% of the elderly woman would result in a posttest probability of approximately 45% (or 0.45). Applying this LR (0.7) to the pretest probability of 0.1% of the young man results in a posttest probability of <0.1%. It becomes evident from these latter two hypothetical examples that the test result has not altered the posttest probabilities to a large extent and further testing is necessary or that the clinician needs to make a decision on the basis of posttest probabilities that are similar to the pretest probabilities. These strategies will differ between the two patients. Most clinicians would remain worried about the elderly women but would safely discharge the young man.

Readers who are interested in the formula for converting pretest probabilities to posttest probabilities will note that it is based on Bayes theorem:

$$\text{posttest odds} = \text{pretest odds} \times \text{likelihood ratio.}$$

Mathematically, we can write this formula as

$$O(D|R) = (O(D) \times P(R|D))/P(R|\bar{D}),$$

where P is the probability of a specific test result R given the status of disease D = disease present, $\bar{D}$ = disease absent, and $O(D)$ is the odds of disease to be calculated as $P(D)/[1 - P(D)]$.

Example 47.1. (Continued)
Returning to our examples, for our elderly female patient with a test result of 1.0 μg/l and pretest odds $= 0.5/(1 - 0.5) = 1$, this formula translates to

$$\text{posttest odds of myocardial infarction} = 1 \times 0.15/0.06 = 2.5.$$

Posttest odds can be converted into posttest probabilities using the following formula:

$$\text{posttest probability} = \text{posttest odds/(posttest odds + 1)} = 2.5/(2.5 + 1) = 71.4\%.$$

This estimate is similar to the estimate on the Fagan Nomogram. In fact, had we been able to use the nomogram as accurately as the calculator we would have obtained identical numbers. This probability moves us into the range of probability where most clinicians would treat patients without further testing because of the morbidity and mortality associated with myocardial infarction.

Example 47.1. (Continued)
For the young male with a troponin of 1.0 μg/l, this formula translates to

$$\text{posttest odds of myocardial infarction} = \text{pretest odds} \times 2.5.$$

We can derive the pretest odds from the pretest probability of 0.1% = 0.001/(1 − 0.001) which is about 0.001.
Thus, it follows that the posttest odds can be calculated as

$$\text{posttest odds} = 0.001 \times 2.5 = 0.0025$$

and the posttest probability as

$$\text{posttest probability} = 0.0025/(0.0025 + 1)$$

which is approximately 0.25%.

Again, the estimate of 0.25% corresponds to the estimate using the Fagan Nomogram but is the exact result of the application of Bayes theorem. Even a troponin I test associated with an LR greater than 1 and commonly considered elevated in the young man does not alter the probability for a myocardial infarction to a great extent.

47.4.4 How to Obtain Pretest Probabilities

We guessed at the pretest probability of the two patients with chest discomfort. How do clinicians obtain valid pretest probabilities? Intuitively, clinicians use their experience based on previous patients with similar presentations. However, the probabilities clinicians and in particular learners assume are prone to bias and error (Richardson 2002; Richardson et al. 2003).

Richardson has suggested that there are two different forms of clinical research that can guide clinicians' estimates of pretest probabilities that need to be distinguished (Richardson 2002). The first type are studies that yield disease probability based on representative patient cohorts with a defined clinical problem that carry out careful diagnostic evaluations and apply explicit diagnostic criteria. Examples include studies on causes of syncope (Soteriades et al. 2002) and cancer in involuntary weight loss (Hernandez et al. 2003). The study by Hill et al. (2004) that provided the estimates for the test properties of troponin I to obtain pretest probabilities could also serve as example. Ideally these prognostic studies are summarized in systematic reviews and evaluated for the confidence that one has in the actual estimate of that pretest probability. The second type of studies is clinical decision rules. These studies assemble cohorts suspected of having the target disorder, apply standard reference tests, and report the frequency of diagnoses in subgroups with identifying clinical features (McGinn et al. 2003). High-quality studies that provide valid estimates of frequency are applicable to our patients and can provide precise estimates of pretest probability. For example, Richardson et al. (2003) estimated in a consecutive patient series that for 78% (95% confidence interval (CI) = (66%, 96%)) of clinical problems evidence of pretest probabilities existed in the literature.

47.4.5 Sensitivity and Specificity

Likelihood ratios help users understand diagnostic tests. However, clinicians use two other descriptive terms for diagnostic tests. It is, therefore, helpful for those with interest in clinical epidemiology to understand these two other terms: sensitivity and specificity.

Table 47.3 Sensitivity and specificity of diagnostic tests ($a + b + c + d = 100\%$)

	Disease or reference standard (proportion)	
	Present (positive)	Absent (negative)
Test results		
Disease present (positive)	True positive (a)	False positive (b)
Disease absent (negative)	False negative (c)	True negative (d)
Sensitivity =	True positive/positive reference standard ***a*/(*a* + *c*)**	
Specificity =		True negative/negative reference standard ***d*/(*b* + *d*)**
Likelihood ratio for positive test (LR+) =	Sensitivity/(1 − Specificity) = $(a/(a+c))/(1-d/(b+d))$	
Likelihood ratio for negative test (LR−) =	(1 − Sensitivity)/Specificity = $(1-(a/(a+c))/(d/(b+d))$	

We could have described the test properties of the study by Hill and colleagues using the concepts of sensitivity and specificity defining normal and abnormal test results. We presented the four different interpretations of troponin I levels, each with the associated likelihood ratios. That classification allowed us to omit the terms normal and abnormal or positive and negative. However, this is not the way most investigators present their result. Investigators also often rely on concepts of sensitivity and specificity.

Sensitivity expresses the proportion of people with the target disorder in whom the test result is positive and specificity expresses the proportion of people without the target disorder in whom a test result is negative. Table 47.3 shows the general concept of sensitivity and specificity in a 2 × 2 table. We could transform the 4 × 2 table described above (Fig. 47.2) into three 2 × 2 tables, depending on what we call positive or negative. Let us assume that only troponin I values $\geq 10.0\,\mu\text{g}/\text{l}$ are positive (or abnormal).

To calculate sensitivity from the data in Table 47.2 for positive troponin I levels, we look at the number of people with proven myocardial infarction ($n = 20$) who were diagnosed as having the target disorder on troponin testing ($n = 3$) showing a sensitivity of 3/20, or approximately 15% ($a/(a + c)$). To calculate specificity, we use the number of people without the target disorder (332) whose troponin test results were classified as normal or $<0.5\,\mu\text{g}/\text{l}$ (303), yielding a specificity of 303/332, or 91% ($d/(b + d)$).

As indicated above, one can easily calculate LRs for different levels of quantitative test results, while sensitivity and specificity require a definition of normal and abnormal that is often arbitrary. In using sensitivity and specificity, one has to either discard important information or recalculate sensitivity and specificity for every cut-point. Therefore, the likelihood ratio is much simpler and much more efficient when tests have more than two possible results, which is very often the case (Guyatt et al. 1990, 1992, 2008a).

47.5 Therapy/Prevention

Example 47.1. (Continued)

Returning to the clinical scenario and having treated the elderly woman with aspirin, clopidogrel, and fondaparinux as well as other relevant medication such as an ACE inhibitor and her other diabetes medication, the team questions whether or not clopidogrel should be continued long term. The fourth-year medical student on the team questions what the evidence related to this treatment is. The resident points you to a study that aimed at maximizing platelet inhibition in patients with acute coronary syndrome, the Clopidogrel in Unstable Angina to Prevent Recurrent Events (CURE) trial (CURE-Investigators 2001). You remember this study and that it is the only one you are aware of addressing this question in patients with acute coronary syndrome. You know that neurologists and cardiologists in your institution often use two antiplatelet agents to maximize platelet inhibition, but you are aware that this leads to an increased bleeding risk and wonder about the quality of the evidence supporting this conclusion and the magnitude of benefits and downsides. An electronic database search confirms that there is no additional evidence addressing this specific question in a randomized trial in patients with acute coronary syndrome (Keller et al. 2007; Vandvik et al. 2012). The CURE investigators addressed the following question. “In patients with acute coronary syndromes without ST-segment elevation, does early and long-term use of clopidogrel plus aspirin versus aspirin alone prevent cardiovascular events and is the combination safe?” You find that this question would be highly relevant to your patient in whom you want to prevent further cardiovascular events.

You decide to critically appraise this study together with your team. The resident retrieves the article through your library’s online full text journal subscription and you evaluate the article with your team over lunch break.

The concepts of clinical epidemiology help clinicians appraise clinical research. We list the three commonly agreed on steps of critical appraisal for studies on therapy or prevention in Table 47.4 and describe how clinical epidemiology facilitates interpretation and evaluation of clinical research. This form addresses critical appraisal for most clinicians. Other important issues for experienced readers concern the appropriateness of the statistical methods.

The first factor that can influence confidence in research results is systematic error or bias. Bias is directly linked to the design and execution of a study. Therefore, the first step is an appraisal of whether results are valid and to what extent bias is present (Table 47.4 – Are the results valid?). The next step in the evaluation of research is the review of the results. Because clinicians often are unfamiliar with applying the magnitude of effects to patient care and because classical epidemiology applies research results in different contexts using different terminology, clinical epidemiologists can bring quantitative outcome measures closer to the clinician. This helps address the question “How large is the effect of an intervention and how large is the role of random error or chance? (What are the results?)” Finally, clinicians need to know whether the results are applicable to their patients. Therefore, clinicians need to appraise whether the results help with decision-making in individual practice circumstances (How can I apply the results to my patient care?). Clinicians must decide whether their patients are similar to those included in the studies from which they obtain the relevant evidence, but clinical epidemiologists can help them in the decision-making process (see Sect. 47.8).

Table 47.4 Critical appraisal of studies about therapy

Question	Therapy or prevention
Study design and execution – evaluation of bias	I. Are the results of the study valid? 1. Were patients randomized? 2. Was randomization concealed? 3. Were patients analyzed in the groups to which they were randomized? 4. Were patients in the treatment and control groups similar with respect to known prognostic factors? 5. Were patients aware of group allocation? 6. Were clinicians aware of group allocation? 7. Were those who collected the data aware of group allocation? 8. Were outcome assessors aware of group allocation? 9. Were data analysts aware of group allocation? 10. Was follow-up complete?
Results and random error	II. What are the results? 1. How large was the treatment effect? 2. How precise was the estimate of the treatment effect?
Application and uptake	III. How can I apply the results to my patient care? 1. Were the study patients similar to my patients? 2. Were all clinically important outcomes considered? 3. Are the likely treatment benefits worth the potential harms and costs?

Critical appraisal (Table 47.4) of the CURE trial reveals that patients were randomized using a 24-h computerized randomization service to conceal randomization. Control patients received placebos and investigators and outcome assessors were blinded to treatment assignment. While blinding refers to not being aware of treatment allocation when treatment has been assigned, concealment refers to avoiding biased allocation of patients because of prior knowledge of forthcoming treatment allocation. Investigators can achieve concealment through measures such as central (telephone) randomization and sealed envelopes. The more stringent the method for concealment, the less likely are those allocating patients to temper with this important aspect of randomized controlled trials. Fulfilling these validity criteria protects against bias, because systematic reviews suggest that lack of blinding and concealment (see Table 47.4) may lead to systematic overestimation of treatment effects although the effects may differ by clinical specialty only (Moher et al. 1998; Balk et al. 2002).

The CURE investigators achieved a follow-up of greater than 99.9% (only 13 out of 12,562 patients were lost) and analyzed patients in the group they were assigned according to the intention to treat principle. The intention to treat principle refers to analysis of patient outcomes based on which group they were randomized regardless of whether they actually received the planned intervention. This analysis preserves the power of randomization, thus maintaining that important unknown factors that influence outcome are likely equally distributed in each comparison group.

Table 47.5 Measures of association and effect

General 2 × 2 table

	Outcome		Absolute risk of outcome:
	+	−	
Intervention (Y)	a	b	Intervention $= a/(a+b) = Y$
Control (X)	c	d	Control $= c/(c+d) = X$

2 × 2 Table for the example from CURE et al. (CURE-Investigators 2001)

	Combined primary outcome		Absolute risk of outcome:
	+	−	
Clopidogrel	1,035	5,224	Clopidogrel = 1,035/(1,035 + 5,224) = 16.5%
Placebo	1,187	5,116	Placebo = 1,187/(1,187 + 5,116) = 18.8%

Absolute risk reduction (*ARR*)
Definition: The difference in risk between the control group and the intervention group
$ARR = c/(c+d) - a/(a+b) = X - Y$
Example:
$ARR = 1{,}187/(1{,}187 + 5{,}116) - 1{,}035/(1{,}035 + 5{,}224) = 2.3\%$ points

Relative risk or risk ratio (*RR*)
Definition: The ratio of risk in the intervention (Y) to the risk in the control group (X)
$RR = Y/X$
Example:
$RR = (1{,}035/(1{,}035 + 5{,}224))/(1{,}187/(1{,}187 + 5{,}116)) = 0.87$

Relative risk reduction (*RRR*)
Definition: The percent reduction in risk in the intervention compared to the control group
$RRR = 1 - RR = (1 - X/Y) \times 100\%$ or
$RRR = [(X - Y)/X] \times 100\%$
Example:
$RRR = (1 - 0.87) \times 100\% = 13\%$

Number needed to treat (*NNT*) – to benefit (NNT_B) or harm (NNT_H)
Definition: Inverse of the *ARR*
$NNT_B = 1/ARR = 1/(X - Y)$
Example:
$NNT_B = 1/2.3\% = 44$

The evaluation of clinical research results requires an understanding of measures of association or effect. As noted above, clinical epidemiologists often use terms that are different from those of classical epidemiologists (see relative risk reduction and number needed to treat below in this section). Table 47.5 summarizes the measures we will now describe in more detail (cf. chapter ▶Rates, Risks, Measures of Association and Impact of this handbook).

47.5.1 Absolute Risk

The easiest measure of risk to understand in clinical epidemiology is the absolute risk. In the CURE trial, the main outcomes were a composite of death from cardiovascular causes, non-fatal myocardial infarction (MI), or stroke and a composite of death from cardiovascular causes, non-fatal MI, stroke, or refractory ischemia. Safety outcomes included major and minor bleeding. We will focus on the latter composite endpoint. The absolute risk for this combined primary endpoint was 16.5%, and the absolute risk for this outcome in the control group was 18.8% (Table 47.5). As described above, other terminologies for the risk of an adverse outcome in the control group are baseline risk, absolute risk, or control event rate.

47.5.2 Absolute Risk Reduction

One can express treatment effects as the difference between the absolute risks in the experimental and control groups, the absolute risk reduction, or the risk difference (*RD*). This effect measure represents the proportion of patients spared from the unfavorable outcome if they receive the experimental therapy (clopidogrel), rather than the control therapy (placebo). In our example, the absolute risk reduction is $18.8\% - 16.5\% = 2.3\%$ points.

47.5.3 Relative Risk

The relative risk or risk ratio presents the proportion of the baseline risk in the control group that still is present when patients receive the experimental treatment (clopidogrel). The relative risk of the combined outcome after receiving clopidogrel is $1{,}035/(1{,}035 + 5{,}224)$ divided by $1{,}187/(1{,}187 + 5{,}116)$ (the risk in the control group), or 0.87. One could also say the risk of experiencing the combined outcome with clopidogrel and aspirin is approximately 87% of that with aspirin alone.

47.5.4 Relative Risk Reduction

The most commonly reported measure of dichotomous treatment effects is the complement of this relative risk, the relative risk reduction. One can obtain the relative risk reduction easily from the relative risk because it is the proportion of baseline risk that is removed by the experimental therapy and it is equivalent to 1.0 – relative risk. It can be expressed as a percent: $(1 - \text{relative risk}) \times 100\% = (1 - 0.87) \times 100\% = 13\%$ or $100\% - 87\% = 13\%$ for this example. Alternatively, one may obtain the relative risk reduction by dividing the absolute risk reduction by the absolute risk in the control group. Therefore, the result is the same if it is calculated from 2.3% (the absolute risk reduction) divided by 18.8% (the risk in the control group) $= 0.13$ (13%). A relative risk reduction of 13% means that clopidogrel reduced the risk of combined outcome by 13% relative to that occurring among control patients. The greater the relative risk reduction, the more efficacious is the therapy. Investigators may compute the relative risk over a period of time,

Table 47.6 Relation between risks and odds

Risk (%)	Risk (proportion)	Odds
80	0.80	4.0000
60	0.60	1.5000
50	0.50	1.0000
40	0.40	0.6667
33	0.33	0.5000
25	0.25	0.3333
20	0.20	0.2500
10	0.10	0.1111
5	0.05	0.0526
1	0.01	0.0101

as in a survival analysis, and call it a hazard ratio, the weighted relative risk over the entire study (see chapter ▶Survival Analysis of this handbook).

In fact, the CURE investigators calculated the hazard ratio which was slightly more in favor of clopidogrel (0.86). For practical purposes, we use the relative risk for the CURE trial in our example. If we had used the hazard ratio for the calculation of the *RR*, the *RRR* would have been 14%.

47.5.5 Odds Ratio

Instead of evaluating the risk of an event, one can estimate the odds of having an event compared with not having an event. Most individuals are familiar with odds in the context of sporting events, when sport reporters describe the odds of a team or player winning a particular event. When odds are used in the medical sciences, it stands for the proportion of patients with the target outcome divided by the proportion without the target outcome. The odds in the control group of the example trial described are 1,187 of 6,303 divided by 5,116 of 6,303. Because the denominator is the same in both the numerator and the denominator, it is canceled out, leaving the number of patients with the event (1,187) divided by the number of patients without the event (5,116). The odds are 1,187/5,116 or 0.232. To convert from odds to risk, one divides the odds by 1 plus the odds. As the odds of the combined endpoint are 0.232, the risk is $0.232/(1 + 0.232)$, or 0.188 (18.8%), identical to the baseline risk reported in the CURE trial. Table 47.6 presents the link between risk and odds. The greater the risk, the greater is the divergence between the risk and odds. Odds and risk are about equal if the absolute risk is small.

In the CURE trial, the odds of the combined endpoint in the clopidogrel group are 1,035 (those with the outcome) compared with 5,224 (those without the outcome), or $1{,}035/5{,}224 = 0.198$, and the odds of the combined endpoint in the placebo group are 0.232. Therefore, the ratio of these odds is (1,035/5,224)/(1,187/5,116), or 0.854. If one used a terminology parallel to risk (note that epidemiologists call a ratio of risks in most instances a relative risk), one would call the ratio of odds relative odds. The commonly used term, however, is odds ratio (*OR*). The odds ratio has been the most popular measure of association. The reason for the use of the odds ratio is that the odds ratio has a statistical advantage because it essentially is independent of the choice between a comparison of the risks of an event (such as death) and the analogous non-event (such as survival), which is not true of the relative risk.

However, clinicians do not easily understand a ratio of odds. Clinicians would like to be able to substitute the relative risk, because it is more intuitively understandable, for the odds ratio. As shown in Table 47.6, as the risk decreases, the odds and risk come closer together. For low event rates, the odds ratio and relative risk are very close. In fact, if the risk is below 25%, odds and risks are approximately equal and many authors calculate relative odds and then report the results as if they calculated relative risks. One can see from the example that the odds ratio of 0.85 is very similar to the relative risk of 0.87. Clinicians should be aware that if events are frequent in either the control group or experimental group, odds ratios can be a very inaccurate estimate of the relative risk. The *RR* and *OR* also will be more similar when the treatment effect is small (*OR* and relative risk are close to 1.0) than when the treatment effect is large. When considering *RR*, *HR*, and *OR*, the *RR* is always closest to unity; the odds ratio is farthest away; and the *HR* is intermediate (Symons and Moore 2002). These differences can become large when effect sizes increase. When event rates are high and effect sizes are large, there are ways of converting the odds ratio to relative risk. Fortunately, clinicians will need to do this infrequently. One note of caution is that typical case-control studies do not yield relative risks.

47.5.6 The Number Needed to Treat and Natural Frequencies

Having seen that making a distinction between odds ratio and relative risk rarely will be important when evaluating clinical research, because high event rates are rare, one must give much more attention to distinguishing between the odds ratio and relative risk compared with the absolute risk reduction. Let us assume that the absolute risk of experiencing the combined outcome would be twice as high in both groups in the CURE trial. This could have happened if the investigators had conducted a study in patients at greater risk for the endpoint, for example, by restricting the study population to older patients. In the clopidogrel group, the absolute risk would become 33% compared with 37.6% in the control group. Therefore, the absolute risk reduction would increase from 2.3 to 4.6%, whereas the relative risk (and therefore the relative risk reduction) would remain identical at 33% divided by 37.6% = 0.87 (the relative risk reduction remains 13%). Therefore, the increase in the proportion of those experiencing the endpoint in both groups by a factor of 2 leaves the relative risk (and the relative risk reduction) unchanged but increases the absolute risk reduction by a factor of 2.

A 13% reduction in the relative risk of the combined endpoint may not sound very impressive; however, its impact on patient groups and practice may be large.

This notion is shown using the concept of the number needed to treat, the number of patients who must receive an intervention during a specific period to prevent one additional adverse outcome or produce one positive outcome (Laupacis et al. 1988). When discussing the number needed to treat, it is important to specify the treatment, its duration, and the outcome being prevented. The number needed to treat is the inverse of the absolute risk reduction, calculated as 1/absolute risk reduction. Therefore, in the example above with an absolute risk reduction of 4.6%, the number needed to treat to benefit (*NNT* or NNT_B) would be 22 (1/4.6%) and it would be 44 (1/2.3%) in the CURE trial. Finally, imagine young patients with no additional risk factors for adverse outcomes. Such patients may carry a baseline risk of 4% for experiencing the endpoint and, therefore, the number needed to treat could increase to 192 $[1/(0.13 \times 4\%)]$. Given the duration, potential harms and cost of treatment with clopidogrel, and the increased risk for bleeding, it could be reasonable to withhold therapy in that latter patient. The *NNT* should always, to prevent misinterpretation, be provided with the time frame that is required to achieve the effect. Unfortunately, we do not know for certain how best to present risk information to patients (Akl et al. 2011). However, presenting the relative risk reduction alone is more persuasive for making actual or hypothetical medical decisions, because the actual benefit appears larger (Bucher et al. 1994; Edwards and Elwyn 1999; Edwards et al. 1999; Akl et al. 2011). Thus, direct to patient advertising and pharmaceutical industry detailing uses relative risk reduction as measure of effect to persuade clinicians and patients to use drug interventions. However, omitting the presentation of baseline risk or not informing those who receive this information that the baseline risk drives the absolute benefit is misleading (Akl et al. 2011). A recent systematic review also showed that there is no clear evidence that the presentation as *NNT* or *ARR* is superior (Akl et al. 2011). Evidence is mounting although not entirely consistent that natural frequencies, numbers presented per 100 or 1,000 where a 33% absolute risk reduction would be expressed as 33 per 100, are better understood and lead to more correct answers (Akl et al. 2011; Woloshin and Schwartz 2011).

47.5.7 How Clinicians Can Use Confidence Intervals

Until now we presented the results of the CURE trial as if they represented the true effect. The results of any experiment, however, represent only an estimate of the truth. The true effect of treatment actually may be somewhat smaller, or larger, than what researchers found. The confidence interval (CI) tells, within the bounds of plausibility, how much smaller or greater the true effect is likely to be. For each of the measures described, one can use statistical programs to calculate confidence intervals.

The point estimate within the confidence interval and the confidence interval itself help with two questions. First, what is the one value most likely to represent the true difference between treatment and control; and second, given the difference between treatment and control, what is the plausible range of differences within which the true effect might actually lie? The smaller the sample size or the number of events in an experiment, the wider the confidence interval. As the sample size gets very large and the number of events increases, investigators become increasingly certain that the truth is not far from the point estimate and, therefore, the confidence interval is narrower.

One can interpret the 95% confidence interval as "what is the range of values of probabilities within which, 95% of the time, the truth would lie?" If investigators or clinicians would not need to be so certain, one could ask about the range within which the true value would lie 90% of the time. This 90% confidence interval would be somewhat narrower.

How does the confidence interval facilitate interpretation of the results from the CURE trial? As described above, the confidence interval represents the range of values within the truth plausibly lies. Accordingly, one way to use confidence intervals is to look at the boundary of the interval that represents the lowest plausible treatment effect and decide whether the action or recommendation would change compared with when one assumes the point estimate represents the truth. Based on the numbers provided in the CURE trial, one can calculate a confidence interval around the point estimate of the relative risk reduction of 13% ranging from approximately 6 to 21%. Values progressively farther from 13% will be less and less likely. One can conclude that patients receiving clopidogrel are less likely to experience the combined endpoint – but the magnitude of the difference may be either quite small (and not outweigh the increased risk of bleeding) or quite large. This way of understanding the results avoids the yes/no dichotomy of testing a hypothesis. Because the lower limit of the CI is associated with a benefit, certainty about beneficial treatment effects from clopidogrel on the combined endpoint is relatively high. However, toxicity and expense will bear on the final treatment decision.

The chief toxicity of clopidogrel the authors of the CURE trial were concerned about was bleeding. They found that the absolute risk increase (*ARI* – conceptually similar to the absolute risk reduction but indicating an increase in risk from the investigational therapy) for major bleeding was 1% point (an increase from 2.7% in the placebo group to 3.7% in the clopidogrel group). The investigators defined major bleeding as substantially disabling bleeding, intraocular bleeding or the loss of vision, or bleeding necessitating the transfusion of at least 2 units of blood. Similarly to the number needed to treat, we can calculate the number of patients who must receive an intervention during a specific period to cause one additional harmful outcome. The number needed to treat to inflict harm (*NNH* or NNT_H) for major bleeding in the CURE trial is equal to 1/*ARI* or $1/0.01 = 100$. The authors also provided the information for minor bleeding which they defined as hemorrhages that led to interruption of the study medication but did not qualify as major bleeding. The risk for minor bleeding was 5.1% in the clopidogrel group compared to 2.4% in the placebo group. Thus, the *NNH* for minor bleeding was $1/0.027 = 37$. Clinicians should keep in mind that estimates of harm also come with uncertainty and, therefore, authors usually present confidence intervals around these harmful effects.

47.5.8 Use of Composite Endpoints

Investigators often use a composition of endpoints when they compare interventions. For example, the CURE trial investigators used a composite endpoint of death from cardiovascular causes, non-fatal MI, stroke, or refractory ischemia. Safety outcomes included major and minor bleeding with the definitions described in the foregoing paragraph. There are a number of reasons for combining several endpoints. First, investigators may believe that distinguishing between outcomes is unnecessary because the endpoints have the same consequences. For example, many investigators combine the outcome ischemic stroke with hemorrhagic stroke because they believe they may be similarly disabling (Lubsen and Kirwan 2002). Second, investigators may choose composite endpoints to avoid misleading conclusion when an intervention reduced an endpoint (usually less severe) by increasing another endpoint (usually more severe) that precludes patients from suffering the less severe endpoint. For example, surgery for cerebrovascular disease could reduce strokes by directly killing those at highest risk for stroke (those who die at surgery are no longer at risk for strokes) (van Walraven et al. 2002). Thus, the use of stroke alone as the endpoint would be misleading. To avoid an erroneous conclusion, an investigator facing this situation must combine the more and less serious outcomes, creating an endpoint such as “stroke or death.” Third, investigators chose to combine endpoints because of the reduced sample size when event rates increase, as one would expect for a combined endpoint. The latter is probably the reason why the CURE investigators used a combined efficacy endpoint.

Using composite endpoints has a number of implications. The most obvious implication is that if these endpoints have different importance to patient (i.e., patients have different underlying values and preferences for these outcomes), treating them as equally important is an oversimplification.

Using combined endpoints does not always support reaching conclusive results in clinical trials. Freemantle et al. (2003) reported on the use of composite endpoints in 167 trials published between 1997 and 2001 in 9 leading medical journals. The authors found that in 69 of these trials the difference between treatment arms in the CEP was not statistically significant. Investigators sometimes included component endpoints that reduced trial efficiency by diluting the treatment effect. For example, Freemantle and colleagues described that the CAPRICORN (Carvedilol Post-Infarct Survival Control in LV Dysfunction) trial investigated the effects of carvedilol, a ß-blocker, in 1,959 patients with left ventricular dysfunction following myocardial infarction (The CAPRICORN Investigators 2001). Originally, the CAPRICORN investigators identified all-cause mortality as primary outcome in the trial protocol. However, while the study was ongoing, the data and safety monitoring board (whose assignment is to protect the patients in a trial) noted that the overall rate of mortality was lower than that predicted for the power analysis and sample size calculation. The board informed the CAPRICORN steering committee of the trial’s insufficient power to identify the primary endpoint as significant (a preset level of significance α of 0.05). Taking the uncommon measure of altering the primary outcome, the steering committee defined a new composite outcome (all-cause mortality or cardiovascular hospital admissions). The steering committee assigned a critical level of significance of $\alpha = 0.045$ to this new composite outcome that they introduced while the trial was ongoing and reduced the significance level of the original primary outcome to achieve statistical significance to $\alpha = 0.005$. They reduced the level of α of the original primary outcome to penalize their retrospective action, but the board decided that if the p-value for either primary outcome achieved statistical significance at the new critical level, the study would be deemed positive.

At the end of the study, the original primary endpoint (all-cause mortality) achieved a p-value of 0.03 (i.e., substantially larger than the new 0.005 allocated after consultation from the data safety and monitoring board but smaller than the original critical level of significance), but the alternative primary outcome achieved a p-value of 0.30. Thus, the original primary outcome did not reach statistical significance at the new and more stringent level and neither did the new composite outcome. Although 12% of patients died in the carvedilol group, compared with 15% in the placebo group, 23% of patients in the carvedilol group and 22% of patients in the placebo group qualified for the composite outcome on the basis of hospitalizations alone, a result that undermined the relatively small reduction in mortality in the carvedilol group. Thus, CAPRICORN provides a neutral result, although the study would have been modestly statistically significant had the original primary outcome of all-cause mortality been maintained.

Of even more concern, composite endpoints can be misleading when the components vary in importance and the least important endpoints dominate the composite (Montori et al. 2005). The use of composites has become widespread, particularly in cardiovascular trials, and trials in which the least important component dominates the results are frequent (Ferreira-Gonzalez et al. 2007).

In summary, evaluating trials that use composite outcome requires scrutiny in regard to the underlying reasons for combining endpoints and its implications and has impact on medical decision-making (see below in Sect. 47.8). Composite endpoints are credible only when the components are of similar importance and the relative effects of the intervention are similar across components (Guyatt et al. 2008a).

47.6 Systematic Reviews

The patient in our scenario took aspirin at a dose of 325 mg on admission. Aspirin use is associated with gastrointestinal bleeding and the risk for bleeding increases with the aspirin dose used (Weil et al. 1995; García Rodríguez et al. 2001). On the other hand, the beneficial effects of lower doses of aspirin on cardiovascular events likely do not differ from those of higher doses (Antithrombotic Trialists' Collaboration 2002). Taken together, studies comparing higher and lower doses of aspirin show similar effects. However, a clinician would ask the following question: "Which of the available studies should I trust and consider for decision-making?" Even for clinicians trained in critical appraisal, evaluating all available studies would be a time-intensive solution.

Traditionally, clinicians – when they did not invest the time and resources to review individual studies – have relied on review articles by authorities in the field. However, experts may be unsystematic in their approach to summarizing the evidence. Unsystematic approaches to identification and collection of evidence risks biased ascertainment. That is, treatment effects may be underestimated or, more commonly, overestimated, and side effects may be exaggerated or ignored. Even if the evidence has been identified and collected in a systematic fashion, if reviewers are then unsystematic in the way they summarize the collected evidence, they run similar risks of bias. In one study, self-rated expertise was inversely related to the methodological rigor of the review (Oxman and Guyatt 1993). One result of unsystematic approaches may be recommendations advocating harmful treatment; in other cases, there may be a failure to encourage effective therapy. For example, experts supported routine use of lidocaine for patients with acute myocardial infarction when available data suggested the intervention was ineffective and possibly even harmful, and they failed to recommend thrombolytic agents for the treatment of acute myocardial infarction when data showed patient benefit (Antman et al. 1992).

Systematic reviews deal with this problem by explicitly stating inclusion and exclusion criteria for evidence to be considered, conducting a comprehensive search for the evidence, and summarizing the results according to explicit rules that include examining how effects may vary in different patient subgroups. When a systematic review pools data across studies to provide a quantitative estimate of overall treatment effect, we call this summary a meta-analysis (cf. chapter ▶Meta-Analysis in Epidemiology of this handbook). Systematic reviews provide higher-quality evidence when the quality of the primary study design is high and other criteria are fulfilled (see section below on GRADE); they typically provide weaker evidence when study designs are observational in nature or other reasons to lower our confidence in the estimates of effect exist, for example, when sample sizes are small and/or designs are poor. Because judgment is involved in many steps in a systematic review (including specifying inclusion and exclusion criteria, applying these criteria to potentially eligible studies, evaluating the methodological quality of the primary studies, and selecting an approach to data analysis), the results of systematic reviews are not immune to bias, for example, publication bias (Atkins et al. 2004; Schünemann et al. 2011a, b).

Nevertheless, in their rigorous approach to identifying and summarizing data, systematic reviews reduce the likelihood of bias in estimating the causal links between management options and patient outcomes.

Over the past 20 to 25 years, the literature describing the methods used in systematic reviews, including studies that provide an empiric basis for guiding decisions about the methods used in summarizing evidence, has rapidly expanded. Clinical epidemiologists have contributed significantly to this development (Egger et al. 2000).

Table 47.7 demonstrates the process of conducting systematic reviews.

As we described above for answering questions in clinical practice (see also Table 47.1), investigators who conduct a systematic review should begin by formulating a clinical question. This question formulation constitutes the essential specific selection criteria for deciding which studies to include in a review. These criteria define the population, the exposures or interventions, the comparison

Table 47.7 The process of conducting systematic reviews

Define the question

- Specify inclusion and exclusion criteria
  - Population
  - Intervention or exposure (and comparison)
  - Outcome
  - Methodology
- Establish a priori hypotheses to explain heterogeneity
- Consider the content of the Summary of Findings (SoF) Table

Conduct literature search

- Decide on information resources: databases, experts, funding agencies, pharmaceutical companies, hand-searching, personal files, registries, citation lists, or retrieved articles
- Determine restrictions: time frame, unpublished data, language
- Identify titles and abstracts

Apply inclusion and exclusion criteria

- Apply inclusion and exclusion criteria to titles and abstracts
- Obtain full articles for eligible titles and abstracts
- Apply inclusion and exclusion criteria to full articles
- Select final eligible articles
- Assess agreement on study selection

Create data abstraction

- Data abstraction: participants, interventions, comparison interventions, study design
- Results
- Methodological quality
- Assess agreement on validity assessment between data abstractors

Conduct analysis

- Determine method for pooling results
- Pool results (if appropriate)
- Decide on handling of missing data

Assess quality of a body of evidence and present results

- Assess the quality of a body of evidence by outcome with the GRADE approach
- Present data in a Summary of Findings (SoF) Table
- Interpret data

intervention, and the outcomes of interest. When formulating the clinical question, authors of systematic reviews should consider all outcomes that are critical or important for decision-making. A good way to present findings from a systematic review is a *Summary of Findings (SoF) Table* which we will describe below (Schünemann et al. 2011a, b). A systematic review will also restrict the included studies to those that meet minimal methodological standards. For example, systematic reviews that address a question of therapy will often include only randomized controlled trials.

Having evaluated the potential eligibility of titles and abstracts, and obtained the full text of potentially eligible studies, reviewers apply the selection criteria to the complete reports. Having completed the data collection process, they assess the methodological quality of the eligible articles and abstract data from each study. Finally, they summarize the data, including, if appropriate, a quantitative synthesis or meta-analysis. The analysis includes an examination of differences among the included studies, an attempt to explain differences in results (exploring heterogeneity), a summary of the overall results, and an assessment of their precision and validity. Guidelines for assessing the validity of reviews and using the results correspond to this process and are available (Oxman et al. 2002).

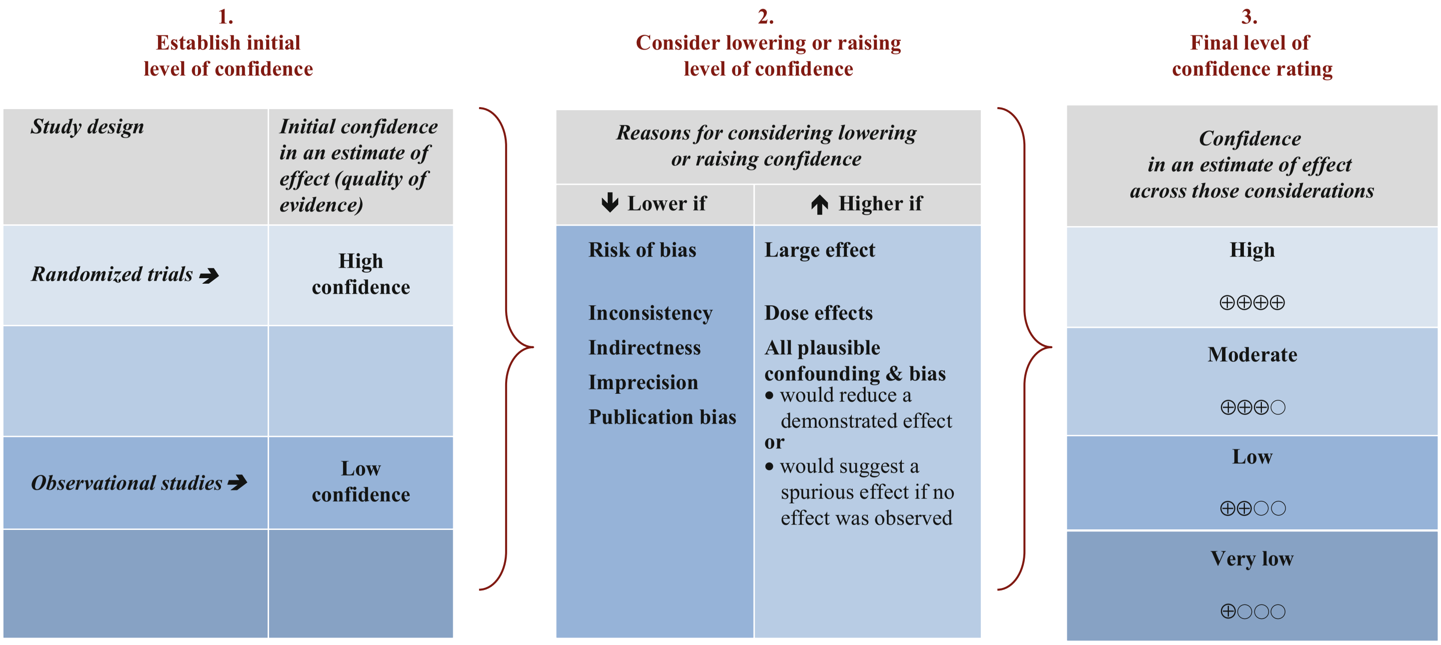
Systematic reviews and other summaries of the evidence should express the confidence users of the review can place in estimates of effect. The GRADE working group has developed an approach to doing this. The next section will focus on this approach and describe a Summary of Findings Table (Guyatt et al. 2011a).

47.7 The GRADE Approach

The GRADE approach assesses the confidence in estimates of effect (i.e., the quality of a body of evidence) for an outcome following a stepwise and transparent process. GRADE classifies the confidence in an estimate of effect into four categories from very low (⊕OOO), low (⊕⊕OO), moderate (⊕⊕⊕O), to high (⊕⊕⊕⊕). The quality of a body of the available best evidence is assessed, beginning with an evaluation of the underlying study design (randomized trials vs. observational studies). This is followed by a detailed assessment of limitations in the design and execution of the studies (risk of bias), imprecision, inconsistency, indirectness, publication bias, magnitude of effect, dose-effect relations, and an assessment of the effect of plausible residual opposing confounding and bias. The overall quality of evidence is determined by the lowest quality of evidence for each of the critical outcomes (Fig. 47.6).

When outcomes point in the same direction (all critical outcomes suggesting benefit), then the overall quality of evidence reflects the quality of the better evidence (e.g., two critical outcomes showing convincing benefit are of low quality and a third of very low quality, the overall quality is not reduced from low to very low). This approach is used by over 70 international organizations, including the World Health Organization (WHO), the National Institute of Health and Clinical Excellence in the UK (NICE), and professional societies such as the American College of Physicians.

One of the outcomes of an assessment following the GRADE approach is a GRADE evidence profile or a Summary of Findings (SoF) Table (for an example, see Fig. 47.7). An SoF Table presents a summary of the key findings from a systematic review in a table, facilitating understandings and identification of key information. There are five sections of the Summary of Findings Table. The first section describes the question the systematic review addresses, including the intervention that was evaluated, to what it was compared and in whom. Next, the table includes information about the desirable and undesirable outcomes that are important to patients and other decision-makers; information for up to seven outcomes can be provided. The third section describes the numerical results for each outcome. In particular, the magnitude of the effect is presented in relative



1.
Establish initial
level of confidence

Study design	*Initial confidence in an estimate of effect (quality of evidence)*
Randomized trials ➔	**High confidence**
Observational studies ➔	**Low confidence**

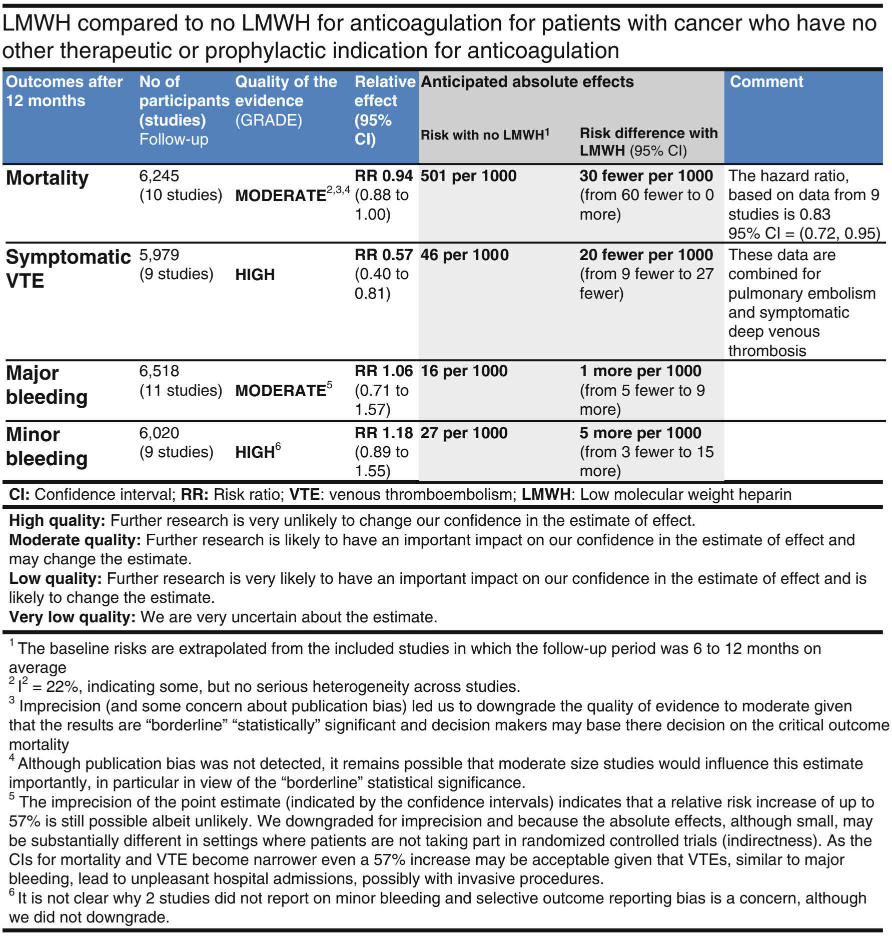
2.
Consider lowering or raising
level of confidence

Reasons for considering lowering or raising confidence	
↓ **Lower if**	↑ **Higher if**
Risk of bias **Inconsistency** **Indirectness** **Imprecision** **Publication bias**	**Large effect** **Dose effects** **All plausible confounding & bias** • would reduce a demonstrated effect **or** • would suggest a spurious effect if no effect was observed

3.
Final level of
confidence rating

Confidence in an estimate of effect across those considerations
High ⊕⊕⊕⊕
Moderate ⊕⊕⊕○
Low ⊕⊕○○
Very low ⊕○○○

Fig. 47.6 GRADE's approach to rating quality of evidence (aka confidence in effect estimates) *For each outcome based on a systematic review and across outcomes (lowest quality across the outcomes critical for decision-making)*



LMWH compared to no LMWH for anticoagulation for patients with cancer who have no other therapeutic or prophylactic indication for anticoagulation

Outcomes after 12 months	No of participants (studies) Follow-up	Quality of the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute effects		Comment
				Risk with no LMWH[1]	Risk difference with LMWH (95% CI)	
Mortality	6,245 (10 studies)	**MODERATE**[2,3,4]	**RR 0.94** (0.88 to 1.00)	**501 per 1000**	**30 fewer per 1000** (from 60 fewer to 0 more)	The hazard ratio, based on data from 9 studies is 0.83 95% CI = (0.72, 0.95)
Symptomatic VTE	5,979 (9 studies)	**HIGH**	**RR 0.57** (0.40 to 0.81)	**46 per 1000**	**20 fewer per 1000** (from 9 fewer to 27 fewer)	These data are combined for pulmonary embolism and symptomatic deep venous thrombosis
Major bleeding	6,518 (11 studies)	**MODERATE**[5]	**RR 1.06** (0.71 to 1.57)	**16 per 1000**	**1 more per 1000** (from 5 fewer to 9 more)	
Minor bleeding	6,020 (9 studies)	**HIGH**[6]	**RR 1.18** (0.89 to 1.55)	**27 per 1000**	**5 more per 1000** (from 3 fewer to 15 more)	

CI: Confidence interval; **RR:** Risk ratio; **VTE**: venous thromboembolism; **LMWH**: Low molecular weight heparin

High quality: Further research is very unlikely to change our confidence in the estimate of effect.
Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Very low quality: We are very uncertain about the estimate.

[1] The baseline risks are extrapolated from the included studies in which the follow-up period was 6 to 12 months on average
[2] I^2 = 22%, indicating some, but no serious heterogeneity across studies.
[3] Imprecision (and some concern about publication bias) led us to downgrade the quality of evidence to moderate given that the results are "borderline" "statistically" significant and decision makers may base there decision on the critical outcome mortality
[4] Although publication bias was not detected, it remains possible that moderate size studies would influence this estimate importantly, in particular in view of the "borderline" statistical significance.
[5] The imprecision of the point estimate (indicated by the confidence intervals) indicates that a relative risk increase of up to 57% is still possible albeit unlikely. We downgraded for imprecision and because the absolute effects, although small, may be substantially different in settings where patients are not taking part in randomized controlled trials (indirectness). As the CIs for mortality and VTE become narrower even a 57% increase may be acceptable given that VTEs, similar to major bleeding, lead to unpleasant hospital admissions, possibly with invasive procedures.
[6] It is not clear why 2 studies did not report on minor bleeding and selective outcome reporting bias is a concern, although we did not downgrade.

Fig. 47.7 Summary of Findings (SoF) summarizing the key findings of a systematic review (developed with Dr. E. Akl) (Akl and Schunemann 2012)

effects (taken directly from the results of the review) and in absolute effects (calculated based on baseline risks). The number of studies and the participants for which there were data are also described. The fourth section provides the assessment of the quality of evidence using the GRADE approach. Finally, there is a section for footnotes and comments. The footnotes and comments describe the judgments behind the quality of the evidence and additional information about how to interpret the results (e.g., validity of an outcome measure). GRADE evidence profiles and SoF Tables can be produced with the GRADE profiler software (freely available for non-profit purposes through www.guidelinedevelopment.org) (Brozek et al. 2011).

Returning to our example, the question you want to address could be formulated as follows: "In patients with coronary artery disease, does low-dose aspirin (75 mg daily or less) compared with high-dose aspirin (325 mg daily or more) confer similar mortality benefits?" Keep in mind that gastrointestinal side effects are more likely with higher doses of aspirin.

A prudent clinician will look for a systematic review to answer this question. A systematic review by the Antithrombotic Trialists' Collaboration provides useful information to answer this question (Antithrombotic Trialists' Collaboration 2002). The investigators carefully evaluated 448 trials for inclusion and finally performed a meta-analysis of 195 trials of antiplatelet effects in cardiovascular disease. Overall they observed that aspirin markedly reduced the risk of recurrent events in patients with acute myocardial infarction (odds ratio 0.7) and patients with previous myocardial infarction (odds ratio 0.75). They identified three trials in patients at high risk for cardiovascular events that compared doses of aspirin of $\leq$75 mg daily to higher doses. Higher doses of aspirin conferred no additional benefit in preventing cardiovascular events when compared with lower doses. In fact the point estimate indicated a benefit of lower doses of aspirin (relative odds reduction 8%). However, these results were not statistically significant (95% CI ranging from approximately -12% indicating a slight benefit with higher doses to 28% indicating a benefit with lower doses). When the investigators compared the effects of low-dose aspirin versus placebo and higher-dose aspirin versus placebo, the effects were similar. Although the evidence from direct comparisons is limited and the investigators included additional patient populations, such as patients with stroke, the systematic review and meta-analysis of all available trials provides additional indirect evidence for comparable effects of higher- and low-dose aspirin. Most clinicians would judge the biology for the role of aspirin in coronary artery disease and other cardiovascular disease to be sufficiently similar to have confidence in this extrapolation. Combining the available evidence in a properly conducted systematic review helps the clinician to obtain answers that are valid and based on methodological evaluation of the literature. Chapter ▶Meta-Analysis in Epidemiology of this handbook addresses methodological issues related to meta-analysis.

Having explained some of the benefits of systematic reviews, we will explain how systematic reviews can be used to guide patient care. Available guidance on therapeutic or prevention goes beyond systematic reviews, because factors in addition to the quality of the evidence and treatment effects are important to make recommendations about treatment (Freemantle et al. 1999). However, systematic reviews should be conducted in the process of generating evidence-based guidelines that provide treatment recommendations (Institute of Medicine 2011).

47.8 Guidelines

Guidelines are systematically developed statements, based on the best available evidence, that should assist providers and recipients of health care and other stakeholders to make informed decisions. Recommendations (in guidelines) may relate to clinical interventions, public health activities, or government policies

(World Health Organization 2003). The key elements of each synthesis include the scope of the guidelines, the interventions and practices considered, the major recommendations and the strength of the evidence and recommendations, and the underlying values and preferences (Guyatt et al. 2002a; Guyatt et al. 2004b; Schünemann et al. 2004, Schunemann et al. 2006a, b; Institute of Medicine 2011; Qaseem et al. 2012; Shekelle et al. 2012). A list of guidelines maintained by the US Agency for Healthcare Research and Quality is available at the National Guidelines Clearinghouse (U.S. Department of Health and Human Services 2013) and by the Guidelines International Network (GIN).

Organizations such as professional societies or governing boards, government agencies, and academic or private institutions typically develop guidelines by convening expert panels. Usually, guideline developers will define the topic of a guideline before evaluating the evidence for clinical sensible questions. Dialogue among clinicians, patients, and the prospective users of the guideline contributes to its refinement. Although it is possible to develop guidelines that are broad in scope, it requires considerable time and resources. Before engaging in a guideline development process, management of conflict of interests has become a key step. This can occur in many different ways and suggestions have been made by several authors and organizations (Schunemann et al. 2009; Guyatt et al. 2010; Eccles et al. 2012; Norris et al. 2012).

One method of defining and focusing the clinical questions of interest and also identifying the processes for which evidence needs to be collected and assessed is the construction of models or causal pathways (Woolf 1994). The causal pathway is a diagram showing the relation of the population, intervention(s) of interest, and the intermediate, surrogate, or definitive health outcomes (Shekelle et al. 1999). When designing the pathway, guideline developers should describe explicitly which outcomes (benefits and harms) they consider important and the associated values. This process reveals specific questions that the evidence must address and where high-quality evidence is lacking, identifying areas for additional research.

While investigators have not tested alternative approaches to guideline development, one suggestion is to include all groups whose activities would be covered by the guidelines and any others with legitimate reasons for having input (Shekelle et al. 1999). A group size of 6 to 15 individuals with clearly identified roles, including group leader and members, specialist resource, technical support, and administrative support, may be advisable (Shekelle et al. 1999). The group should comprise members proficient in the following areas: literature searching and retrieval, epidemiology, biostatistics, health services research, clinical area of interest (generalists and specialists), group process, writing, and editing.

Over the past few years, several groups have suggested standards for the preparation of clinical practice guidelines (Shiffman et al. 2003; Schünemann et al. 2004; Schunemann et al. 2006a, b; Institute of Medicine 2011; Qaseem et al. 2012). A high-quality clinical practice guideline should consider the following steps which include a systematic review of the literature:

1. Define explicit criteria to search for evidence. Similar to the clinical sensible questions identified above for searching the evidence and conducting systematic reviews, this step should include a clear definition of the population, the

intervention and comparison intervention, and the outcome of interest. The development of a clinical pathway helps in identifying the components of the clinical question and in identifying gaps.

2. Define explicit eligibility criteria for the identified evidence. Guideline development groups should define eligibility criteria for the evidence that they wish to include. Examples include restricting evidence to randomized controlled trials or studies that have used validated instruments for functional outcomes or assessed mortality.
3. Conduct or use comprehensive searches for evidence. A guideline development group should ensure that they conduct a complete evaluation of the evidence. The group may either use a high-quality systematic review developed by others or conduct their own systematic review for each recommendation they make in their guideline.
4. Perform a standard consideration of study quality. If a systematic review is identified that answers the group's clinical question and may suffice for developing the guideline, the group should still consider an evaluation of the individual study quality. Should the group conduct a systematic review or update an existing review, the evaluation of study quality becomes an essential step in conducting a systematic review. Quality evaluation includes looking at the basic study design, the detailed study design and execution, and the directness of evidence. The directness of the evidence refers to reporting of surrogate outcomes, for example, deep venous thrombosis by ultrasonography as risk factor for fatal events, versus outcomes such as mortality.
5. Summarize the evidence. Guideline development groups who use high-quality systematic reviews often will find meta-analysis of studies included in the systematic review. The summaries help in obtaining estimates of intervention effects. It is important to note that summary estimates should be available for all important outcomes, both beneficial and harmful. Ideally the summary estimates also would be available for cost.
6. Acknowledge values and preferences underlying the group's recommendations (Schunemann et al. 2006a, b). Many guideline reports take for granted that guideline developers adequately represent patients' interests. The latter is not necessarily correct and there is a risk that, for example, a specialty society may recommend procedures where the benefits may not outweigh the risks or costs (Woolf et al. 1999). For example, the American Urologic Society and the American Cancer Society recommend prostate cancer screening with prostate-specific antigen (PSA) for men older than 50, while the American College of Preventive Medicine and the US Preventive Services Task Force (USPSTF) do not make this recommendation (Ferrini and Woolf 1998; American Urological Association 2000; USPSTF 2002; Smith et al. 2003). Thus, there should be clear statements about which principles, such as patient autonomy, non-maleficence, or distributive justice, were given priority in guiding decisions about the value of alternative interventions to inform users of the guidelines (Shekelle et al. 1999). Guidelines should report whether it is intended to optimize values for individual patients, reimbursement agencies, or society as a whole. Groups that

ensure representation by experts in research methodology, practicing generalists and specialists, and public representatives are more likely to have considered diverse views in their discussions than groups limited to content area experts.

7. Grade the strength of recommendations. Because clinicians are interested in the strength of a recommendation and the balance of benefits and risks, the next section is devoted to recommendations and the grading of the strength of recommendations. Many organizations use the GRADE approach to doing this.

Recommendations Treatment decisions involve balancing likely benefits against harms and costs. Evidence-based guidelines and treatment recommendations are systematic syntheses of the best available evidence that provide clinicians with guidance for treating average patients in clinical practice. To integrate recommendations with their own clinical judgment, clinicians need to understand the basis for the clinical recommendations that experts offer them. A common systematic approach to grading the strength of treatment recommendations can minimize bias and aid interpretation.

As part of the first American College of Chest Physician (ACCP) Consensus Conference on Antithrombotic Treatment in 1986, Sackett suggested a formal rating scheme, derived from the Canadian Task Force on the Periodic Health Examination, for assessing levels of evidence (Canadian Task Force on the Periodic Health Examinsation 1979; Sackett 1986). During the past 25 years, clinical epidemiologists have lead the evolution of these "rules of evidence" (Cook et al. 1992; Guyatt et al. 1995, 2001, 2011a, b, 2002a, b; Atkins et al. 2004; Brozek et al. 2011), which experts have applied to generate grades of recommendations.

The strength of any recommendation depends on four factors: the trade-off between benefits and downsides and the quality of the body of evidence that leads to estimates of the treatment effect, values and preferences, and resource considerations. The GRADE approach to grading the quality of evidence and strength of recommendations captures the key issues of guideline development. GRADE generates strong or weak (aka conditional) recommendations either in favor or against a management strategy based on evaluation of the four factors that influence the strength of recommendations (Table 47.8).

Examples of recommendations developed with the GRADE approach are now ubiquitous in the medical literature (Schunemann et al. 2007a, b; Brozek et al. 2010; Fiocchi et al. 2010; Falzon et al. 2011; Qaseem et al. 2011; Akl et al. 2012; Bates et al. 2012; Santesso et al. 2012). Many guideline developers use frameworks such as the one shown in Table 47.9 to enhance transparency and record the process and reasons for recommendations (Santesso et al. 2012).

Table 47.8 Reasons and definition of strong and weak recommendations according to GRADE

Strength	Definition
Strong	The guideline panel is confident that the desirable effects of adherence to the recommendation outweigh the undesirable effects
Conditional	The guideline panel concludes that the desirable effects of adherence to a recommendation probably outweigh the undesirable effects

Table 47.9 GRADE approach to moving from evidence to recommendations: enhancing transparency when moving from evidence to recommendations

Question/recommendation:

Population/patients:

Intervention:

Setting (if relevant):

Decision domain	**Explanation**	**Summary of reason for judgment**	**Judgment**
Quality of evidence (QoE) *Is there high or moderate quality evidence?* The higher the quality of evidence, the more likely is a strong recommendation	**Consider** **QoE for benefits:** **QoE for harms:** **QoE for resource use:** • *Key reasons for down- or upgrading?* • *All critical outcomes measured?*		Yes No □ □
Balance of benefits versus harms and burdens *Are you confident that the benefits outweigh the harms and burden or vice versa?* The larger the difference between the benefits and harms and the certainty around that difference, the more likely is a strong recommendation. The smaller the net benefit or net harm and the lower the certainty for that net effect, the more likely is a conditional/weak recommendation	**Baseline risk for benefits, harms and burden:** • *Is the baseline risk similar across subgroups?* • *Should there be separate recommendations for subgroups?* **Relative risk for benefits and harms:** • *Are the relative benefits large?* • *Are the relative harms large?* **Requirement for modeling:** • *Is there a lot of extrapolation and modeling required for these outcomes?* **Average values:** • *What are the average values?* • *Are there differences in the relative value of the critical outcomes?*		Yes No □ □

Values and preferences *Are you confident about the assumed or identified relative values and are they similar across the target population?* The more certainty or similarity in values and preferences, the more likely a strong recommendation	**Perspective taken:** **Perspective taken:** • *Patients or general population/society?* **Source of values:** **Source of variability if any:** **Method for determining values satisfactory for this recommendation:**	Yes ☐ No ☐
Resource implications *Are the resources worth the expected net benefit from following the recommendation?* The lower the cost of an intervention compared to the alternative and other costs related to the decision – that is, the fewer resources consumed – the more likely is a strong recommendation in favor of that intervention	**What are the costs per resource unit?** **Feasibility:** • *Is this intervention generally available?* **Opportunity cost:** • *Is this intervention and its effects worth withdrawing or not allocating resources from other interventions?* **Differences across settings:** • *Is there lots of variability in resource requirements across settings?*	Yes ☐ No ☐
Overall strength of recommendation		
Remarks		
Evidence to recommendation synthesis		

Returning to our clinical scenario and applying the GRADE approach to grading of recommendation, the use of clopidogrel in addition to aspirin generated the following recommendation regarding the dual use of aspirin and clopidogrel over aspirin alone in the last iteration of the ACCP antithrombotic guidelines (Vandvik et al. 2012):

For patients in the first year after an acute coronary syndrome who have not undergone percutaneous coronary intervention, [the guideline panel] recommends dual antiplatelet therapy [. . .] clopidogrel 75 mg daily plus low-dose aspirin 75–100 mg daily over single antiplatelet therapy (strong recommendation, moderate quality evidence).

In the next section, we will describe the importance of health-related quality of life (HRQL) outcomes followed by a section on integrating patient preferences in decision-making and how clinical epidemiology is key in obtaining patients' preferences and values.

47.9 Health-Related Quality of Life Instruments and Their Application in Clinical Studies

Clinical journals have published trials in which HRQL instruments are the primary outcome measures. With the expanding importance of HRQL in evaluating new therapeutic interventions, investigators (and readers) are faced with a large array of instruments. Researchers have proposed different ways of categorizing these instruments, according to the purpose of their use, into instruments designed for screening, providing health profiles, measuring preference, and making clinical decisions (Osoba 1991), or into discriminative and evaluative instruments.

We have also suggested a taxonomy based on the domains of HRQL which an instrument attempts to cover (Guyatt et al. 1989). According to this taxonomy, an HRQL instrument may be categorized, in a broad sense, as generic or specific. *Generic instruments* cover (or at least aim to cover) the complete spectrum of function, disability, and distress of the patient and are applicable to a variety of populations. Within the framework of generic instruments, health profiles and utility measures provide two distinct approaches to measurement of global quality of life. *Specific instruments* are focused on disease or treatment issues specifically relevant to the question at hand.

47.9.1 Generic Instruments

Health profiles are single instruments that measure multiple different aspects of quality of life. They usually provide a scoring system that allows aggregation of the results into a small number of scores and sometimes into a single score (in which case, it may be referred to as an index). As generic measures, their design allows their use in a wide variety of conditions. For example, one health profile, the Sickness Impact Profile (SIP), contains 12 "categories" which can be aggregated into two dimensions and five independent categories and also into a single overall score (Bergner et al. 1981). The SIP has been used in studies of cardiac rehabilitation (Ott et al. 1983), total hip joint arthroplasty (Liang et al. 1985), and treatment of back pain (Deyo et al. 1986). In addition to the SIP, there are a number of other health profiles available: the Nottingham Health Profile (Hunt et al. 1980), the Duke-UNC Health Profile (Parkerson et al. 1981), and the McMaster Health Index Questionnaire (Sackett et al. 1977). Increasingly, a collection of related instruments from the Medical Outcomes Study (Tarlov et al. 1989) has become the most popular and widely used generic instruments. Particularly popular is one version that includes 36 items, the SF-36 (Brook et al. 1979; Ware and Sherbourne 1992; Ware et al. 1995). The SF-36 is available in over 40 languages and normal values for the general population in many countries are available.

While each health profile attempts to measure all important aspects of HRQL, they may slice the HRQL pie quite differently. For example, the McMaster Health Index Questionnaire follows the World Health Organization approach and identifies three dimensions: physical, emotional, and social. The Sickness Impact Profile includes a physical dimension (with categories of ambulation, mobility, body care, and movement), a psychosocial dimension (with categories including social interaction and emotional behavior), and five independent categories including eating, work, home management, sleep and rest, and recreations and pastimes.

General health profiles offer a number of advantages to clinical investigators. Their reproducibility and validity have been established, often in a variety of populations. When using them for discriminative purposes, one can examine and establish areas of dysfunction affecting a particular population. Identification of these areas of dysfunction may guide investigators who are constructing disease-specific instruments to target areas of potentially greatest impact on the quality of life. Health profiles, used as evaluative instruments, allow determination of the effects of an intervention on different aspects of quality of life, without necessitating the use of multiple instruments (and thus saving both the investigator's and the patient's time). Because health profiles are designed for a wide variety of conditions, one can potentially compare the effects on HRQL of different interventions in different diseases. Profiles that provide a single score can be used in a cost-effectiveness analysis, in which the cost of an intervention in dollars is related to its outcome in natural units.

The main limitation of health profiles is that they may not focus adequately on the aspects of quality of life specifically influenced by a particular intervention. This may result in an inability of the instrument to detect a real effect in the area of importance (i.e., lack of responsiveness). In fact, disease-specific instruments offer greater responsiveness compared with generic instruments (Guyatt et al. 1999; Wiebe et al. 2003; Brozek et al. 2009). We will return to this issue when we discuss the alternative approach, specific instruments.

47.9.2 Specific Instruments

An alternative approach to HRQL measurement is to focus on aspects of health status that are specific to the area of primary interest. The rationale for this approach lies in the increased responsiveness that may result from including only those aspects of HRQL that are relevant and important in a particular disease process or even in a particular patient situation. One could also focus an instrument only on the areas that are likely to be affected by a particular drug.

In other situations, the instrument may be specific to the disease (instruments for chronic lung disease, for rheumatoid arthritis, for cardiovascular diseases, for endocrine problems, etc.); specific to a population of patients (instruments designed to measure the HRQL of the frail elderly, who are afflicted with a wide variety of different diseases); specific to a certain function (questionnaires which examine emotional or sexual function); or specific to a given condition or problem (such as pain) which can be caused by a variety of underlying pathologies. Within a single condition, the instrument may differ depending on the intervention. For example, while success of a disease-modifying agent in rheumatoid arthritis should result in improved HRQL by enabling a patient to increase performance of physically stressful activities of daily living, occupational therapy may achieve improved HRQL by encouraging family members to take over activities formerly accomplished with difficulty by the patient. Appropriate disease-specific HRQL outcome measures should reflect this difference.

Specific instruments can be constructed to reflect the "single state" (how tired have you been: very tired, somewhat tired, full of energy) or a "transition" (how has your tiredness been: better, the same, worse) (MacKenzie and Charlson 1986). The same could be said of generic instruments. Specific measures can integrate aspects of morbidity, including events such as recurrent myocardial infarction (Olsson et al. 1986).

The disease-specific instruments may be used for discriminative purposes. They may aid, for example, in evaluating the extent to which a primary symptom (e.g., dyspnea) is related to the magnitude of physiological abnormality (e.g., exercise capacity) (Mahler et al. 1987). Disease-specific instruments can be applied for evaluative purposes to establish the impact of an intervention on a specific area of dysfunction, and hence aid in elucidating the mechanisms of drug action (Jaeschke et al. 1991). Guidelines provide structured approaches for constructing specific measures (Guyatt et al. 1986). Whatever approaches one takes to the construction of disease-specific measures, a number of head-to-head comparisons between generic and specific instruments suggest that the latter approach will fulfill its promise of enhancing an instrument's responsiveness, the ability to detect change in HRQL (Tandon et al. 1989; Tugwell et al. 1990; Chang et al. 1991; Laupacis et al. 1991; Smith et al. 1993; Goldstein et al. 1994).

In addition to the improved responsiveness, specific measures have the advantage of relating closely to areas routinely explored by the physician. For example, a disease-specific measure of quality of life in chronic lung disease focuses on dyspnea during day-to-day activities, fatigue, and areas of emotional dysfunction, including frustration and impatience (Guyatt et al. 1987). Specific measures may therefore appear clinically sensible to the clinician.

The disadvantages of specific measures are that they are (deliberately) not comprehensive and cannot be used to compare across conditions or, at times, even across programs. This suggests that there is no one group of instruments that will achieve all the potential goals of HRQL measurement. Thus, investigators may choose to use multiple instruments. Some of these instruments are preferences or value instruments that can also be used for clinical decision-making described in the next section.

47.10 Integrating Patient Preferences in the Decision-Making Process and Resolution of the Clinical Scenario

In reviewing the data from the CURE trial, you conclude that if the patient before you does not receive clopidogrel as additional therapy, the best estimate of the risk for recurrent myocardial infarction or any of the other endpoints summarized in the composite endpoint in the trial during the next year is 16.5% and, further, that clopidogrel is likely to decrease this risk by approximately 13%, corresponding to absolute risk reductions (*ARR*) of 2.3% points over a 1-year period. As described above, this translates into a number needed to treat (*NNT*) for 1 year to prevent one of the endpoints summarized in the composite endpoint of approximately 44 for treatment with clopidogrel (Table 47.5).

Examining the likelihood of major bleeding, the CURE trial suggests an absolute risk increase of 1.0%. This estimate translates into a *NNH* of 100. In light of your knowledge that the patient before you is intelligent, conscientious, and very concerned about his health, you anticipate a high rate of adherence; in addition, you anticipate that the bleeding risk rate of 1% (or the *NNT* of 100) represents a good estimate for risk of the patient in front of you.

Considering these numbers, you are aware that the treatment decision may depend on the relative value the patient places on avoiding a recurrent myocardial infarction or any of the other endpoints summarized in the composite endpoint in the CURE trial and avoiding a major bleeding including those leading to vision loss. We have pointed out that since there are always advantages and disadvantages to an intervention, evidence alone cannot determine the best course of action. Patients, their proxies, or, if a parental approach to decision-making is desirable, the clinician as decision-maker must always trade the benefits, harm, and costs associated with alternative treatment strategies, and values and preferences always bear on those trade-offs. Findings that assume that the patient's values are similar to one's own, yet this may be incorrect. For example, facing a decision concerning anticoagulation in atrial fibrillation, clinicians are more concerned about bleeding risk, and place less weight on the associated stroke reduction, than patients (Devereaux et al. 2001). Thus, a fundamental principle of EBM is the explicit inclusion of patients and society's values and clinical circumstances in the clinical decision-making process as described before in Fig. 47.1b (Haynes et al. 2002a, b). Clinical epidemiology can help identifying and applying the key issues in the decision-making process.

Considering the model by Haynes et al. (2002a, b) and our updated model, you are now faced with the problem of how to best incorporate the patient's values into the decision. Before resolving the scenario, we will describe different ways to optimize decision-making.

For many – perhaps most – of our clinical decisions, the trade-off is sufficiently clear that clinicians need not concern themselves with variability in patient values. Previously healthy patients will all want antibiotics to treat their pneumonia or their urinary tract infection, anticoagulation to treat their pulmonary embolus, or aspirin to treat their myocardial infarction. Under such circumstances, a brief explanation of the rationale for treatment and the expected benefits and side effects will suffice.

When benefits and risks are balanced more delicately and the best choice may differ across patients, clinicians must attend to the variability in patients' values (such as in a weak recommendation in the GRADE approach to grading recommendations). One fundamental strategy for integrating evidence with preferences involves communicating the benefits and risks to patients, thus permitting them to incorporate their own values and preferences in the decision. One advantage of this approach is that it avoids the vexing problem of measuring patients' values. Unfortunately, the problem of communicating the evidence to patients in a way that allows patients to clearly and unequivocally understand their choices is almost as vexing as the direct measurement of patient values.

A second basic strategy is to ascertain the relative value patients place on the key outcomes associated with the management options. One can then consider the likely outcomes of alternative courses of action and use the patient's values as the basis of trading off benefits and risks. When done in a fully quantitative way, this approach becomes a decision analysis using individual patient preferences (Guyatt et al. 2002a, b). A number of texts provide information on decision analysis and decision analyses are available for a number of topics, such as the prevention of ischemic stroke with warfarin in atrial fibrillation (Petitti 1994; Thomson et al. 2000; Guyatt et al. 2002a, b).

In addition, patients often have preferences not only about the outcomes, but about the decision-making process itself. These preferences can vary, and the patient's desired level of involvement should determine which approach the clinician takes (Strull et al. 1984; Degner et al. 1997; Stiggelbout and Kiebert 1997). Ethicists have characterized the alternative strategies (Emanuel and Emanuel 1992). At one end of the spectrum, the physician acts as a technician, providing the patient with information and taking no active part in the decision-making process. This corresponds to the first strategy for incorporating patient values, presenting patients with the likely benefits, risks, inconvenience, and cost and then letting patients decide. At the opposite extreme, corresponding to the second strategy, ascertaining the patient's values and then making a recommendation in light of the likely advantages and disadvantages of alternative management approaches, the clinician takes a "paternalistic" approach and decides what is best for the patient in light of that patient's preferences.

However, intermediate approaches of shared decision-making are generally more popular than those at either extreme. Shared decision-making uses both of the two fundamental approaches to decision-making presented above: the clinician typically shares the evidence, in some form, with the patient while simultaneously attempting to understand the patient's values. Evidence that more active patient involvement in the process of health care delivery can improve outcomes and reported quality of life – and, possibly, reduce health care expenditures – provides empirical evidence in support of secular trends toward patient autonomy and away from paternalistic approaches (Greenfield et al. 1988; Stewart 1995; Szabo et al. 1997).

Clinicians should temper their enthusiasm for active patient involvement in decision-making with an awareness that many patients prefer paternalistic approaches. For example, the results of a survey of 2,472 patients suffering from chronic disease (hypertension, diabetes, heart failure, myocardial infarction, or depression) completed between 1986 and 1990 supported this approach (Arora and McHorney 2000). In response to the statement: "I prefer to leave decisions about my medical care up to my doctor," 17.1% strongly agreed, 45.5% agreed, 11.1% were uncertain, 22.5% disagreed, and only 4.8% strongly disagreed. In another study of node-negative breast cancer patients considering adjuvant chemotherapy, 84% of 171 women preferred an independent or shared role in decision-making (Whelan et al. 2003). Increasing general levels of education, the advent of the Internet and the resulting access to medical information, and an increasingly litigious and consumerist environment have all contributed to patients wishing to play a more active role in decision-making and may explain a shift in patient preferences for decision-making. Shared decision-making and patient-centeredness have become attractive approaches to resolving the profusion of challenging choices facing patients and clinicians (Charles et al. 1999a, b; Edwards and Elwyn 2001; Guyatt et al. 2004a).

Regardless of the decision-making approach chosen by the patient and clinician, integrating values and preferences and communicating options inject challenges into the process by insisting that clinicians consider quantitative estimates of benefits and risks, rather than just whether a treatment works or whether toxicity occurs. If clinicians leave the decisions to patients, they must effectively communicate the probabilities associated with the alternative outcomes to them. If they opt for taking responsibility for combining patient values with the evidence, they must quantify those values. A vague sense of the patient's preferences cannot fully satisfy the rigor of the optimal decision-making approach.

We will now describe some of the specific strategies associated with two decision-making models: one in which the clinician presents the patient with the likely consequences of alternative management strategies and leaves the choice to the patient, and the other in which the clinician ascertains the patient's values and provides a recommendation.

47.10.1 Patient as Decision-Maker: Decision Aids

If the patient wishes to play the primary role in decision-making, clinicians may use intuitive approaches to communicating concepts of risk and risk reduction that they have developed through clinical experience. They will answer the patient's questions and ultimately act on the patient's decision. Alternatively, if available for a particular decision, clinicians can use a decision aid that presents descriptive and probabilistic information about the disease, treatment options, and potential outcomes in a patient-friendly manner (Levine et al. 1992; Holmes-Rovner et al. 2001; O'Connor 2001; Barry 2002; Akl et al. 2007; Breslin et al. 2008; Mullan et al. 2009; Montori et al. 2011).

A well-constructed decision aid has two advantages. One is that someone has reviewed the literature and produced a rigorous summary of the probabilities. Clinicians who doubt that the summary of probabilities is rigorous can go back to the original literature on which those probabilities are based and determine their accuracy. A second advantage of a well-constructed decision aid is that it will offer a pretested and effective way of communicating the information to patients who may have little background in quantitative decision-making. Most commonly, decision aids use visual props to present the outcome data in terms of the percentage of people with a certain condition who do well without intervention, compared to the percentage who do well with intervention. Decision aids will summarize the data regarding all outcomes of importance to patients.

Theoretically, decision aids present an attractive strategy for ensuring that patient values guide clinical decision-making. What impact do decision aids actually have on clinical practice? Stacey and colleagues conducted a systematic review, finding 86 randomized trials that used decision aids, for example, the decision for or against hormone replacement therapy in women after menopause or decisions related to breast surgery in breast cancer (Stacey et al. 2003, 2011).

The findings of this systematic review suggest that decision aids "increase people's involvement and improve knowledge and realistic perception of outcomes" but the size of the effect varies across studies. Decision aids reduce the choice of interventions such as discretionary surgery and have no apparent adverse effects on health outcomes or satisfaction.

47.10.2 Patient as Provider of Values

The second set of approaches all begin with, at minimum, establishing the relative value the patient places on the target outcomes. Doing so requires that the patient understand the nature of those outcomes. How, for instance, would a patient with atrial fibrillation facing a decision about using oral anticoagulation to prevent strokes imagine living with a stroke, or the experience of having a gastrointestinal bleeding episode as a side effect of the oral anticoagulation? Patients may find

Table 47.10 Sample descriptions of major stroke and gastrointestinal bleed

Major stroke	Bleeding
• You suddenly are dizzy and blackout • You are unable to move one arm and one leg • You cannot swallow or control bladder and bowel • You are unable to understand what is being said • You are unable to talk • You feel no physical pain • You are admitted to hospital • You cannot dress • The nurse feeds you • You cannot walk • After 1 month with physiotherapy, you are able to wiggle your toes and lift your arm off the bed • You remain this way for the rest of your life • Another illness will likely cause your death	• You feel unwell for 2 days then suddenly you vomit blood • You are admitted to hospital • You stop taking warfarin • A doctor puts a tube down your throat to see where you are bleeding from • You receive sedation to ease the discomfort of the test • You do not need an operation • You receive blood transfusions to replace the blood you lost • You stay in hospital 1 week • You feel well at the end of your hospital stay • You need to take pills for the next 6 months to prevent further bleeding • You do not take warfarin any more • After that you are back to normal

a written description of the health states (Table 47.10) useful in the process of describing their preferences (Devereaux et al. 2001).

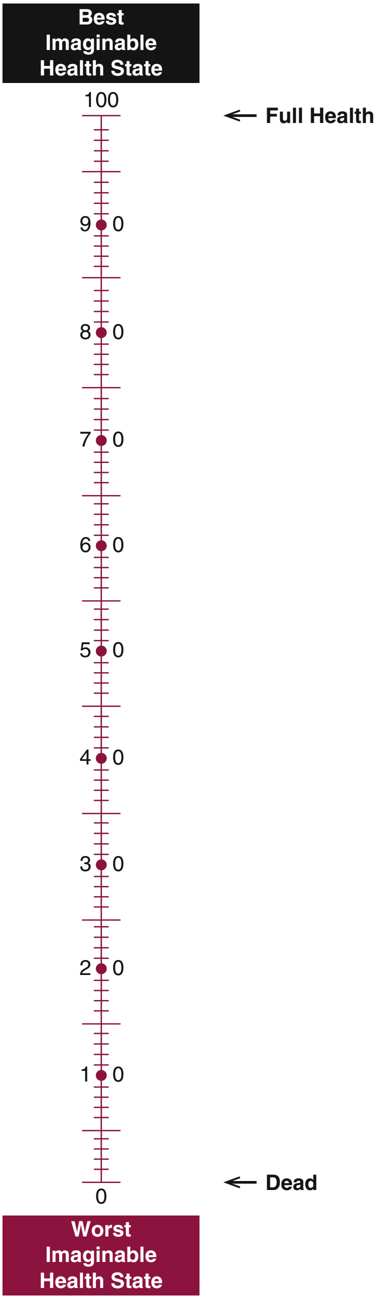
Having made their best effort to ensure that patients understand the outcomes, clinicians can choose from among a number of ways of obtaining their values for those outcomes. They can gain a qualitative sense of their patients' preferences from a discussion without a formal structure. Alternatively, a direct comparison between outcomes may prove useful. For instance, with only two outcomes, the patient can make a direct comparative rating. The question may be as follows: "How much worse would it be to have a stroke versus a gastrointestinal bleeding episode? Would it be equally bad? Twice as bad to have a stroke? Three times as bad?"

Using a somewhat more complex strategy, the clinician can ask the patient to place a mark on a visual analogue scale or "feeling thermometer," in which the extremes are anchored at dead and full health, to represent how the patient feels about the health states in question (Fig. 47.8).

When, as in the case of a gastrointestinal bleeding and a stroke, some health states are temporary and others are permanent, the clinician must ensure that patients incorporate the duration of the health state in their rating.

More sophisticated approaches include the time trade-off and the standard gamble (Torrance 1986). In completing the time trade-off, patients choose between a longer period in a state of impaired health (such as recovery from severe stroke) and a shorter period in a state of full health. With the standard gamble, by contrast, patients are asked to choose between living in a state of impaired health versus taking a gamble in which they may return to full health or die immediately. These latter approaches may come much closer to meeting assumptions that health

Fig. 47.8 The feeling thermometer



economists argue are necessary for accurate ratings of the relative value of health states in the context of choice with uncertain outcomes.

Regardless of the strategy clinicians use to obtain patient values, they must somehow integrate these values with the likely outcomes of the alternative management strategies. Formal decision analysis provides the most rigorous method for making this integration. Practical software for plugging in the patient's values and conducting a patient-specific decision analysis for common clinical problems is being developed, although not yet available for routine use in daily clinical practice. Investigators have shown that, when patients' values are used in individualized decision analyses, their decisions about anticoagulation in atrial fibrillation differ from those suggested by existing guidelines (Protheroe et al. 2000). Whether the decisions would have differed had the patients been provided with the probabilities and asked to choose their preferred management strategy – as with a decision aid – remains unknown.

Even if the tools for individual decision analysis were widely available, application of the approach would depend on the availability of clinicians who could devote time to eliciting patient values. Such a process may be resource intensive, and issues of how much gain there is from the investment, or the intervention's cost-effectiveness, may become very important. Exactly the same considerations apply to the use of decision aids, in which the improvement of knowledge is clear but the impact on anxiety, or on the choices patients actually make, is not as obvious.

Another method of expressing information to patients that incorporates their values is the likelihood of being helped versus harmed (Sackett et al. 2000). Clinicians can apply the likelihood of being helped versus harmed to any clinical decision, and preliminary evidence suggests the approach may be useful on busy clinical services. The clinician begins by calculating the *NNT* and *NNH* for the average patients in a study or studies from which the data about treatment effectiveness and harm come. The clinician then adjusts the average *NNT* and *NNH* for the individual patient according to that patient's likelihood of suffering the target event that treatment is intended to prevent, and the risks it may precipitate, relative to the average patient. Having established the relative likelihood of help versus harm, the clinician explores the patient's values about the severity of adverse events that might be caused by the treatment relative to the severity of the target event that treatment helps prevent. The final adjustment of the likelihood of being helped versus harmed incorporates the patient's values without providing formal help by a decision aid.

47.10.3 Likelihood of Help Versus Harm

For the sake of simplicity, we will assume that the patient in our scenario places a mean value on the composite endpoints cardiovascular disease (defined by the CURE investigators as death from cardiovascular causes, non-fatal MI, stroke, or refractory ischemia), major bleeding (substantially disabling bleeding, intraocular bleeding or the loss of vision, or bleeding necessitating the transfusion of at least 2 units of blood), and minor bleeding (any other bleeding leading to interruption of study medication), respectively. We will ignore other factors bearing on the decision, such as taking an additional pill daily.

During your discussion with the patient about the consequences of further cardiovascular disease and major bleeding, you asked her to use the "feeling thermometer" (see Fig. 47.8) to estimate how she feels about each of the two combined outcomes. We will ignore minor bleeding episodes in this example, because your patient is not concerned at all about the risk and consequences of minor bleeding. However, she places a mean value of suffering additional consequences of cardiovascular disease at 0.2 and of living with a major bleed at 0.7. You use these on your handheld personal digital assistant to calculate her likelihood of being helped or harmed (*LHH*) from clopidogrel therapy versus placebo therapy.

Using the *NNT*s calculated in the scenario, the *LHH* for clopidogrel versus placebo becomes:

$$\begin{aligned} LHH &= (1/NNT):(1/NNH) \\ &= (1/44):(1/100) = 100/44 \\ &= 2.3. \end{aligned}$$

(Note: we could also use (1/absolute risk reduction):(1/absolute risk increase) but this uses decimal fractions and may increase the likelihood of arithmetic errors.)

Therefore, you can tell the patient that clopidogrel is approximately twice as likely to help her as to harm her, when compared with placebo.

Incorporating her values that you elicited, the *LHH* becomes

$$\begin{aligned} LHH &= (1/NNT) \times (1-\mathrm{Uevent}):(1/NNH) \times (1\text{–}\mathrm{Utoxicity}) \\ &= (1/44) \times (1-0.2):(1/100) \times (1-0.7) \\ &= 6.1, \end{aligned}$$

where Uevent is the value of the outcome prevented (composite endpoint of death from cardiovascular causes, non-fatal MI, stroke, or refractory ischemia) and Utoxicity (major bleeding) is the value of the side effect.

You can now inform the patient that clopidogrel is approximately six times as valuable to help her as to harm her. Including additional outcomes would increase the number of terms in the numerator (benefits) or denominator (adverse consequences).

Alternatively, a quicker way of incorporating the patient's values is to ask the patient to rate one event against another. For example, is the adverse effect about as severe as the event the treatment prevents – or ten times as bad or only half as severe? This rating ("*s*") can then be used to adjust the *LHH* as

$$\text{LHH} = (1/\text{NNT}) \times \text{s}:(1/\text{NNH}).$$

Having ascertained the likely outcomes of the alternate courses of action, the clinician must either present patients with the options and outcomes and leave it for them to choose, try to discover the patient's values and having done so suggest a course of action to the patient (the paternalistic approach), or choose the middle course of shared decision-making. The patient's preferred decision-making style will guide the clinician in this regard. However, communicating the nature of the outcomes and their probabilities in a way the patient will understand, or accurately ascertaining the patient's values regarding the outcomes, remains problematic.

The challenges of optimal clinical decision-making should not obscure the realization that clinicians face these challenges in helping patients with every management decision. For each choice, clinicians guide patients with their best estimate of the likely outcomes. They then help patients balance these outcomes in making their ultimate decision. Finding better strategies to carry out these tasks remains a frontier for clinical epidemiology.

47.10.4 Semistructured Conversation and Resolution

Example 47.1. (Continued)

You discuss the option of clopidogrel therapy with your patient who is feeling better now and appears to have a good understanding of the information you are providing. You explain that – based on your assessment and the patients' values and preferences – benefit and harm of clopidogrel are pointing clearly toward more desirable than undesirable consequences: for every 44 patients treated for 1 year in the CURE trial, there was one less occurrence of the combined endpoint. However, for every 100 patients treated with clopidogrel for 1 year, one additional patient suffered a major bleeding episode, and for every 37 patients treated for 1 year, there was one additional minor bleeding episode. You also explain that, because these are only estimates, the true effect might be somewhat smaller or larger for both the benefit and harms. Your patient states she agrees that clopidogrel for the longer terms seems to be a good choice.

Because the decision to take clopidogrel depends to some degree on the patient's values and preferences regarding preventing the combined endpoint versus incurring additional risk of bleeding and you have the results of the calculation of the likelihood of being helped or harmed. You explain that the results of your calculation using the software on your personal digital assistant support her preference for taking clopidogrel. In terms of expense, you are uncertain about the cost of clopidogrel. Because you feel that this question will come up with additional patients and that you had wanted to address it for some time, you call the hospital pharmacist. She informs you that the cost for clopidogrel is approximately $90 per month and that at least one analysis has suggested the drug is not cost-effective (Gaspoz et al. 2002). The patient tells you that she has minimal co-pay for most medications and remains interested in taking the medication. Together, you decide that beginning clopidogrel treatment clearly is in her best interest and you start the patient on a 300 mg loading dose of clopidogrel and continue with 75 mg daily. You also suggest reducing the dose of aspirin as lower doses of aspirin confer similar benefits and doses of 75–325 mg were given in the CURE trial together with clopidogrel. She is reassured by the fact that most patients in her situation choose the same option for therapy.

47.11 Conclusions

We have presented several concepts of clinical epidemiology. Working through the clinical problems has implicitly highlighted some of clinical epidemiology's research challenges. The following makes explicit some of these challenges that clinical epidemiologists and other investigators need to tackle in future research. A number of important areas have been identified and we will list them here briefly. It is not clear what are the best ways to educate clinicians and students in the methodology of clinical epidemiology, and educational researchers will have to focus on this aspect. The Cochrane Collaboration, other organizations, and researchers will further elaborate the methodology of systematic reviews (e.g., of diagnostic studies, observational studies, and health-related quality of life outcomes). Obtaining further information about the most valid and informative ways of presenting statistical information and education of patients and clinicians about these issues is an important task for clinical epidemiologists. Research should also focus on improving the development of clinical practice guidelines and integrating cost information in guidelines and recommendations (Schünemann et al. 2004). Furthermore, research on implementing guidelines into clinical practice is an area of intensive research (e.g., DECIDE 2012). It is clear that studies of guideline implementation should follow the same methodological rigor as other studies, but they are presented with different challenges, such as the need for large cluster randomized clinical trials. We described the integration of preferences and values in medical decision-making as well as bedside decision-making in particular above. Tools that facilitated this difficult task are in development. Health decision aids, in particular electronic decision aids, promise to advance this science. Along with health decision aids, the integration of HRQL information into clinical practice and guidelines presents challenges that investigators need to resolve (Frost et al. 2003). Finally, conducting additional research of integrating electronic health (eHealth) (including multimedia decision aids) into clinical practice presents a fascinating but challenging outlook.

References

Akl EA, Schunemann HJ (2012) Routine heparin for patients with cancer? One answer, more questions. N Engl J Med 366(7):661–662

Akl EA, Grant BJ, Guyatt GH, Montori VM, Schunemann HJ (2007) A decision aid for COPD patients considering inhaled steroid therapy: development and before and after pilot testing. BMC Med Inform Decis Mak 7:12

Akl EA, Oxman AD, Herrin J, Vist GE, Terrenato I, Sperati F, Costiniuk C, Blank D, Schunemann H (2011) Using alternative statistical formats for presenting risks and risk reductions. Cochrane Database Syst Rev Mar 16;(3):CD006776

Akl EA, Kennedy C, Konda K, Caceres CF, Horvath T, Ayala G, Doupe A, Gerbase A, Wiysonge CS, Segura ER, Schünemann HJ, Lo YR (2012) Using GRADE methodology for the development of public health guidelines for the prevention and treatment of HIV and other STIs among men who have sex with men and transgender people. BMC Public Health 12(1):386

American Urological Association (2000) Prostate-specific antigen (PSA) best practice policy. Oncology 14:267–286

Antithrombotic Trialists' Collaboration (2002) Collaborative metaanalysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients. BMJ 324:71–86

Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC (1992) A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. JAMA 268:240–248

Arora N, McHorney C (2000) Patient preferences for medical decision making: who really wants to participate? Med Care 38:335–341

Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D, Hill S, Jaeschke R, Leng G, Liberati A, Magrini N, Mason J, Middleton P, Mrukowicz J, O'Connell D, Oxman AD, Phillips B, Schunemann HJ, Edejer TT, Varonen H, Vist GE, Williams JW Jr, Zaza S (2004) Grading quality of evidence and strength of recommendations. BMJ 328(7454):1490

Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, Lau J (2002) Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials [comment]. JAMA 287(22):2973–2982

Barry MJ (2002) Health decision aids to facilitate shared decision making in office practice. Ann Intern Med 136(2):127–135.

Bates SM, Jaeschke R, Stevens SM, Goodacre S, Wells PS, Stevenson MD, Kearon C, Schunemann HJ, Crowther M, Pauker SG, Makdissi R, Guyatt GH (2012) Diagnosis of DVT: antithrombotic therapy and prevention of thrombosis, 9th ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest 141(2 Suppl):e351S–418S

Bergner M, Bobbitt R, Carter W, Gilson B (1981) The sickness impact profile: development and final revision of a health status measure. Med Care 19:787–805

Bernardo JM, Adrian FM (1994) Bayesian theory. Smith Wiley, Chichester

Berry DA (1996) Statistics: a Bayesian perspective. Wadworth/Duxbury Press, Belmont

Bhandari M, Montori V, Devereaux PJ, Dosanjh S, Sprague S, Guyatt GH (2003) Challenges to the practice of evidence-based medicine during residents' surgical training: a qualitative study using grounded theory. Acad Med 78(11):1183–1190

Breslin M, Mullan RJ, Montori VM (2008) The design of a decision aid about diabetes medications for use during the consultation with patients with type 2 diabetes. Patient Educ Couns 73(3):465–472

Brook R, Ware J, Davies-Avery A, Stewart A, Donald C (1979) Overview of adult health status measures fielded in Rand's health insurance study. Med Care 17(Suppl 7):iii-55

Brozek JL, Guyatt GH, Heels-Ansdell D, Degl'Innocenti A, Armstrong D, Fallone CA, Wiklund I, van Zanten SV, Chiba N, Barkun AN, Akl EA, Schunemann HJ (2009) Specific HRQL instruments and symptom scores were more responsive than preference-based generic instruments in patients with GERD. J Clin Epidemiol 62(1):102–110

Brozek JL, Bousquet J, Baena-Cagnani CE, Bonini S, Canonica GW, Casale TB, van Wijk RG, Ohta K, Zuberbier T, Schunemann HJ (2010) Allergic Rhinitis and its Impact on Asthma (ARIA) guidelines: 2010 revision. J Allergy Clin Immunol 126(3):466–476

Brozek J, Oxman AD, Schünemann HJ (2011) GRADEprofiler (GRADEpro). Version 3.6. McMaster University, Hamilton

Bucher HC, Weinbacher M, Gyr K (1994) Influence of method of reporting study results on decision of physicians to prescribe drugs to lower cholesterol concentration [comment]. BMJ 309(6957):761–764

Canadian Task Force on the Periodic Health Examination (1979) The periodic health examination. Can Med Assoc J 121:1193–1254

Chang S, Fine R, Siegel D, Chesney M, Black D, Hulley S (1991) The impact of diuretic therapy on reported sexual function. Arch Intern Med 151:2402–2408

Charles C, Gafni A, Whelan T (1999a) Decision-making in the physician-patient encounter: revisiting the shared treatment decision-making model. Soc Sci Med 49(5):651–661

Charles C, Whelan T, Gafni A (1999b) What do we mean by partnership in making decisions about treatment? BMJ 319(7212):780–782

Cook DJ, Guyatt GH, Laupacis A, Sackett DL (1992) Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest 102(4_Suppl):305S–311S, Erratum in: Chest 1994, 105:647

CURE-Investigators (2001) Effects of clopidogrel in addition to aspirin in patients with acute coronary syndromes without ST-segment elevation. N Engl J Med 345:494–502

Daly J (2005) Evidence-based medicine and the search for a science of clinical care. Memorial Fund and University of California Press, Berkeley

DECIDE (2012) Developing and evaluating communication strategies to support informed decisions and practice based on evidence. http://www.decide-collaboration.eu. Accessed 5 May 2013

Degner L, Kristjanson L, Bowman D, Sloan JA, Carriere KC, O'Neil J, Bilodeau B, Watson P, Mueller B (1997) Information needs and decisional preferences in women with breast cancer. JAMA 277:1485–1492

Devereaux PJ, Anderson DR, Gardner MJ, Putnam W, Flowerdew GJ, Brownell BF, Nagpal S, Cox JL (2001) Differences between perspectives of physicians and patients on anticoagulation in patients with atrial fibrillation: observational study. BMJ 323(7323):1218–1222

Deyo R, Diehl A, Rosenthal M (1986) How many days of bed rest for acute low back pain? a randomized clinical trial. N Engl J Med 315:1064–1070

Diamond GA (1999) The wizard of odds: Bayes theorem and diagnostic testing [comment]. Mayo Clin Proc 74(11):1179–1182

Diamond GA, Forrester JS (1979) Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. N Engl J Med 300(24):1350–1358

Diamond GA, Forrester JS, Hirsch M, Staniloff HM, Vas R, Berman DS, Swan HJ (1980) Application of conditional probability analysis to the clinical diagnosis of coronary artery disease. J Clin Investig 65(5):1210–1221

Diamond GA, Hirsch M, Forrester JS, Staniloff HM, Vas R, Halpern SW, Swan HJ (1981) Application of information theory to clinical diagnostic testing: the electrocardiographic stress test. Circulation 63:915–921

Dolan JG, Bordley DR, Mushlim AI (1986) An evaluation of clinicians' subjective prior probability estimates. Med Decis Mak 6:216–223

Eccles MP, Grimshaw JM, Shekelle P, Schunemann HJ, Woolf S (2012) Developing clinical practice guidelines: target audiences, identifying topics for guidelines, guideline group composition and functioning and conflicts of interest. Implement Sci 7(1):60

Edwards A, Elwyn G (1999) How should effectiveness of risk communication to aid patients' decisions be judged? A review of the literature [comment]. Med Decis Mak 19(4):428–434

Edwards A, Elwyn G (2001) Evidence-based patient choice. Inevitable or impossible? Oxford University Press, New York

Edwards A, Elwyn G, Stott N (1999) Communicating risk reductions. Researchers should present results with both relative and absolute risks. BMJ 318(7183):603; author reply 603–604

Egger M, Davey Smith G, Altman DG (2000) Meta-analysis in context. Systematic reviews in health care. BMJ Books, London

Emanuel E, Emanuel L (1992) Four models of the physician-patient relationship. JAMA 267: 2221–2226

Evidence-Based-Medicine-Working-Group (1992) Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA 268:2420–2425

Fagan T (1975) Nomogram for Bayes theorem. N Engl J Med 293:257

Falzon D, Jaramillo E, Schunemann HJ, Arentz M, Bauer M, Bayona J, Blanc L, Caminero JA, Daley CL, Duncombe C, Fitzpatrick C, Gebhard A, Getahun H, Henkens M, Holtz TH, Keravec J, Keshavjee S, Khan AJ, Kulier R, Leimane V, Lienhardt C, Lu C, Mariandyshev A, Migliori GB, Mirzayev F, Mitnick CD, Nunn P, Nwagboniwe G, Oxlade O, Palmero D, Pavlinac P, Quelapio MI, Raviglione MC, Rich ML, Royce S, Rusch-Gerdes S, Salakaia A, Sarin R, Sculier D, Varaine F, Vitoria M, Walson JL, Wares F, Weyer K, White RA, Zignol M (2011) WHO guidelines for the programmatic management of drug-resistant tuberculosis: 2011 update. Eur Respir J 38(3):516–528

Feinstein A (1968) Clinical epidemiology I. The population experiments of nature and of man in human illness. Ann Intern Med 69:807–820

Ferreira-Gonzalez I, Busse JW, Heels-Ansdell D, Montori VM, Akl EA, Bryant DM, Alonso-Coello P, Alonso J, Worster A, Upadhye S, Jaeschke R, Schunemann HJ, Permanyer-Miralda G, Pacheco-Huergo V, Domingo-Salvany A, Wu P, Mills EJ, Guyatt GH (2007) Problems with use of composite endpoints in cardiovascular trials: systematic review of randomised controlled trials. BMJ 334(7597):786

Ferrini R, Woolf S (1998) American College of Preventive Medicine Practice Policy: screening for prostate cancer in American men. Am J Prev Med 15:81–84

Fiocchi A, Brozek J, Schunemann H, Bahna SL, von Berg A, Beyer K, Bozzola M, Bradsher J, Compalati E, Ebisawa M, Guzman MA, Li H, Heine RG, Keith P, Lack G, Landi M, Martelli A, Rance F, Sampson H, Stein A, Terracciano L, Vieths S (2010) World Allergy Organization (WAO) Diagnosis and Rationale for Action against Cow's Milk Allergy (DRACMA) Guidelines. Pediatr Allergy Immunol 21(Suppl 21):1–125

Fletcher RH, Fletcher SW, Wagner EH (1996) Clinical epidemiology: the essentials, 3rd edn. Williams & Wilkins, Baltimore

Freemantle N, Mason J, Eccles M (1999) Deriving treatment recommendations from evidence within randomized trials: the role and limitation of meta-analysis. Int J Technol Assess Health Care 15(2):304–315

Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C (2003) Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA 289(19):2554-2559

Frost M, Bonomi A, Cappelleri JC, Schünemann HJ, Moynihan TJ, Aaronson NK, Clinical Significance Consensus Meeting Group (2003) Applying quality of life data formally and systematically into clinical practice. Clin Ther 25:D10

García Rodríguez L, Hernández-Díaz S, de Abajo FJ (2001) Association between aspirin and upper gastrointestinal complications: systematic review of epidemiologic studies. Br J Clin Pharmacol 52:563–571

Gaspoz JM, Coxson PG, Goldman PA, Williams LW, Kuntz KM, Hunink MG, Goldman L (2002) Cost effectiveness of aspirin, clopidogrel, or both for secondary prevention of coronary heart disease. N Engl J Med 346(23):1800–1806

Glass RD (1996) Diagnosis: a brief introduction. Oxford University Press, Melbourne

Goldstein R, Gort E, Guyatt G, Stubbing D, Avendano M (1994) Prospective randomized controlled trial of respiratory rehabilitation. Lancet 344:1394–1397

Greenfield S, Kaplan SH, Ware JE Jr, Yano EM, Frank HJ (1988) Patients' participation in medical care: effects on blood sugar control and quality of life in diabetes. J Gen Intern Med 3(5): 448–457

Grundy SM, Cleeman JI, Merz CN, Brewer HB Jr, Clark LT, Hunninghake DB, Pasternak RC, Smith SC Jr, Stone NJ (2004) Implications of recent clinical trials for the National Cholesterol Education Program Adult Treatment Panel III guidelines. Circulation 110(2):227–239

Guyatt GH (1991) Evidence-based medicine. ACP J Club 114:A-16

Guyatt G (2002a) Introduction. In: Guyatt G, Rennie D (eds) Users' guide to the medical literature: a manual for evidence-based clinical practice. American Medical Association, Chicago, pp 3–13

Guyatt G (2002b) Preface. In: Guyatt G, Rennie D (eds) Users' guide to the medical literature: a manual for evidence-based clinical practice. American Medical Association, Chicago, p xiii

Guyatt G, Bombardier C, Tugwell P (1986) Measuring disease-specific quality-of-life in clinical trials. Can Med Assoc J 134:889–895

Guyatt G, Walter S, Norman G (1987) Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 40:171–178

Guyatt G, Zanten SVV, Feeny D, Patrick D (1989) Measuring quality of life in clinical trials: a taxonomy and review. Can Med Assoc J 140:1441–1447

Guyatt G, Patterson C, Ali M, Singer J, Levine M, Turpie I, Meyer R (1990) Diagnosis of iron-deficiency anemia in the elderly [comment]. Am J Med 88(3):205–209

Guyatt G, Oxman AD, Ali M, Willan A, McIlroy W, Patterson C (1992) Laboratory diagnosis of iron-deficiency anemia: an overview [erratum appears in J Gen Intern Med 1992 Jul-Aug;7(4):423]. J Gen Intern Med 7(2):145–153

Guyatt GH, Sackett D, Sinclair JC, Hayward R, Cook DJ, Cook RJ (1995) Users' guides to the medical literature IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group. JAMA 274:1800–1804

Guyatt G, King DR, Feeny DH, Stubbing D, Goldstein RS (1999) Generic and specific measurement of health-related quality of life in a clinical trial of respiratory rehabilitation. J Clin Epidemiol 52:187–192

Guyatt G, Schünemann H, Cook D, Pauker S, Sinclair J, Bucher H, Jaeschke R (2001) Grades of recommendation for antithrombotic agents. Chest 119(1 Suppl):3S–7S

Guyatt G, Hayward RS, Richardson WS, Green L, Wilson MC, Sinclair J, Cook D, Glasziou P, Detsky A, Bass E (2002a) Moving from evidence to action. In: Guyatt G, Rennie D (eds) Users' guide to the medical literature: a manual for evidence-based clinical practice. American Medical Association, Chicago, pp 175–204

Guyatt G, Sinclair J, Cook D, Jaeschke R, Schünemann H (2002b) Moving from evidence to action. Grading recommendations – a qualitative approach. In: Guyatt GH, Rennie D (eds) Users' guides to the medical literature: a manual for evidence-based clinical practice. American Medical Association, Chicago

Guyatt G, Devereaux P, Montori V, Schünemann H, Bhandari B (2004a) Putting the patient first: in our practice, and in our use of language. Evidence Based Medicine 9:6–7

Guyatt G, Schünemann H, Cook D, Jaeschke R, Pauker S (2004b) Applying the grades of recommendation for antithrombotic and thrombolytic therapy: the seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. Chest 126(3_Suppl):179S–187S

Guyatt G, Rennie D, Meade M, Cook D (2008a) Users' guide to the medical literature: a manual for evidence-based clinical practice. McGraw Hill, Chicago

Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ, for the GRADE Working Group (2008b) GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 336(7650):924–926

Guyatt G, Akl EA, Hirsh J, Kearon C, Crowther M, Gutterman D, Lewis SZ, Nathanson I, Jaeschke R, Schunemann H (2010) The vexing problem of guidelines and conflict of interest: a potential solution. Ann Intern Med 152(11):738–741

Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, DeBeer H, Jaeschke R, Rind D, Meerpohl J, Dahm P, Schunemann HJ (2011a) GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 64(4):383–394

Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A (2011b) GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 64(4):380–382

Haynes RB, Devereaux PJ, Guyatt GH (2002a) Clinical expertise in the era of evidence-based medicine and patient choice. Vox Sang 83(Suppl 1):383–386

Haynes RB, Devereaux PJ, Guyatt GH (2002b) Clinical expertise in the era of evidence-based medicine and patient choice. ACP J Club 136(2):A11–A14

Hernandez JL, Riancho JA, Matorras P, Gonzalez-Macias J (2003) Clinical evaluation for cancer in patients with involuntary weight loss without specific symptoms. Am J Med 114(8):631–637

Hill SA, Devereaux PJ, Griffith L , Opie J, McQueen MJ, Panju A, Stanton E, Guyatt GH (2004) Can troponin I measurement predict short-term serious cardiac outcomes in patients presenting to the emergency department with possible acute coronary syndrome? Can J Emerg Med 6: 22–30

Holmes-Rovner M, Llewellyn-Thomas H, Entwistle V, Coulter A, O'Connor A, Rovner DR (2001) Patient choice modules for summaries of clinical effectiveness: a proposal. BMJ 322(7287):664–667

Hunt S, McKenna S, McEwen J, Backett E, Williams J, Papp E (1980) A quantitative approach to perceived health status: a validation study. J Epidemiol Community Health 34:281–286

Institute of Medicine (2011) Clinical practice guidelines we can trust. The National Academies Press, Washington, DC

Jaeschke R, Singer J, Guyatt G (1991) Using quality-of-life measures to elucidate mechanism of action. Can Med Assoc J 144:35–39

Keller TT, Squizzato A, Middeldorp S (2007) Clopidogrel plus aspirin versus aspirin alone for preventing cardiovascular disease. Cochrane Database Syst Rev Jul 18;(3):CD005158; Review. Update in: Cochrane Database Syst Rev 2011; 1:CD005158

Ladenheim ML, Kotler TS, Pollock BH, Berman DS, Diamond GA (1987) Incremental prognostic power of clinical history, exercise electrocardiography and myocardial perfusion scintigraphy in suspected coronary artery disease. Am J Cardiol 59(4):270–277

Laupacis A, Sackett D, Roberts RS (1988) An assessment of clinically useful measures of the consequences of treatment. N Engl J Med 318:1728–1733

Laupacis A, Wong C, Churchill D (1991) The use of generic and specific quality-of-life measures in hemodialysis patients treated with erythropoietin. Control Clin Trials 12:168S–179S

Ledley RS, Lusted LB (1959) Reasoning foundations of medical diagnosis: symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science 130: 9–21

Levine MN, Gafni A, Markham B, MacFarlane D (1992) A bedside decision instrument to elicit a patient's preference concerning adjuvant chemotherapy for breast cancer [comment]. Ann Intern Med 117(1):53–58

Liang M, Larson M, Cullen K, Schwartz J (1985) Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 28:542–547

Lubsen J, Kirwan B (2002) Combined endpoints: can we use them? Stat Med 21:2959–2970

MacKenzie M, Charlson M (1986) Standards for the use of ordinal scales in clinical trials. BMJ 292:40–43

Mahler D, Rosiello R, Harver A, Lentine T, McGovern J, Daubenspeck J (1987) Comparison of clinical dyspnea ratings and psychophysical measurements of respiratory sensation in obstructive airway disease. Am Rev Respir Dis 135:1229–1233

McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG (2003) Clinical prediction rules. In: Guyatt GH, Rennie D (eds) Users' guide to the medical literature. American Medical Association, Chicago, pp 471–484

McKibbon A, Hunt D, Richardson W, Hayward R, Wilson M, Jaeschke R, Haynes RB, Wyer P, Craig J, Guyatt GH (2002) Finding the evidence. In: Guyatt GH, Rennie D (eds) Users' guides to the medical literature: a manual for evidence-based clinical practice. American Medical Association, Chicago, p 16

Meier MA, Al-Badr WH, Cooper JV, Kline-Rogers EM, Smith DE, Eagle KA, Mehta RH (2002) The new definition of myocardial infarction: diagnostic and prognostic implications in patients with acute coronary syndromes [comment]. Arch Intern Med 162(14):1585–1589

Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP (1998) Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? [comment]. Lancet 352(9128):609–613

Montori VM, Permanyer-Miralda G, Ferreira-Gonzalez I, Busse JW, Pacheco-Huergo V, Bryant D, Alonso J, Akl EA, Domingo-Salvany A, Mills E, Wu P, Schunemann HJ, Jaeschke R, Guyatt GH (2005) Validity of composite endpoints in clinical trials. BMJ 330(7491):594–596

Montori VM, Shah ND, Pencille LJ, Branda ME, Van Houten HK, Swiglo BA, Kesman RL, Tulledge-Scheitel SM, Jaeger TM, Johnson RE, Bartel GA, Melton LJ 3rd, Wermers RA (2011) Use of a decision aid to improve treatment decisions in osteoporosis: the osteoporosis choice randomized trial. Am J Med 124(6):549–556

Mullan RJ, Montori VM, Shah ND, Christianson TJ, Bryant SC, Guyatt GH, Perestelo-Perez LI, Stroebel RJ, Yawn BP, Yapuncich V, Breslin MA, Pencille L, Smith SA (2009) The diabetes mellitus medication choice decision aid: a randomized trial. Arch Intern Med 169(17): 1560–1568

Norris SL, Burda BU, Holmer HK, Ogden LA, Fu R, Bero L, Schunemann H, Deyo R (2012) Author's specialty and conflicts of interest contribute to conflicting guidelines for screening mammography. J Clin Epidemiol 65(7):725–733

O'Connor A (2001) Using patient decision aids to promote evidence-based decision making. ACP J Club 135(1):A11–A12

Olsson G, Lubsen J, Van Es G, Rehnqvist N (1986) Quality-of-life after myocardial infarction: effect of long term metoprolol on mortality and morbidity. BMJ 292:1491–1493

Osoba D (1991) Effect of cancer on quality-of-life. CRC, Boston

Ott C, Sivarajan E, Newton K, Almes MJ, Bruce RA, Bergner M, Gilson BS (1983) A controlled randomized study of early cardiac rehabilitation: the sickness impact profile as an assessment tool. Heart Lung 12:162–170

Oxman A, Guyatt G (1993) The science of reviewing research. Ann N Y Acad Sci 703:125–133. Discussion 133–124

Oxman AD, Sackett DL, Guyatt GH (1993) Users' guides to the medical literature. I. How to get started. The evidence-based medicine working group. JAMA 270(17):2093–2095

Oxman A, Guyatt G, Cook D, Montori V (2002) Summarizing the evidence. In: Guyatt GH, Rennie D (eds) Users' guide to the medical literature: a manual for evidence-based clinical practice. American Medical Association, Chicago, pp 155–174

Parkerson G, Gehlback S, Wagner E, Sherman A, Clapp N, Muhlbaier L (1981) The Duke-UNC Health Profile: an adult health status instrument for primary care. Med Care 19:806–828

Paul J (1938) Clinical epidemiology. J Clin Investig 17:539–541

Petitti D (1994) Meta-analysis, decision analysis and cost-effectiveness analysis. Oxford University Press, Oxford

PIOPED-Investigators (1990) Value of ventilation/perfusion scan in acute pulmonary embolism. Results of the Prospective Investigation of Pulmonary Embolism (PIOPED). JAMA 263: 2753–2759

Protheroe J, Fahey T, Montgomery AA, Peters TJ (2000) The impact of patients' preferences on the treatment of atrial fibrillation: observational study of patient based decision analysis [comment]. BMJ 320(7246):1380–1384

Qaseem A, Wilt TJ, Weinberger SE, Hanania NA, Criner G, van der Molen T, Marciniuk DD, Denberg T, Schunemann H, Wedzicha W, MacDonald R, Shekelle P (2011) Diagnosis and management of stable chronic obstructive pulmonary disease: a clinical practice guideline update from the American College of Physicians, American College of Chest Physicians, American Thoracic Society, and European Respiratory Society. Ann Intern Med 155(3): 179–191

Qaseem A, Forland F, Macbeth F, Ollenschlager G, Phillips S, van der Wees P (2012) Guidelines international network: toward international standards for clinical practice guidelines. Ann Intern Med 156(7):525–531

Richardson WS (2002) Five uneasy pieces about pre-test probability [comment]. J Gen Intern Med 17(11):882–883

Richardson WS, Wilson MC, Nishikawa J, Hayward RS (1995) The well-built clinical question: a key to evidence-based decisions. ACP J Club 123(3):A12–A13

Richardson WS, Wilson MC, Williams JW Jr, Moyer VA, Naylor CD (2003) Clinical manifestation of disease. In: Guyatt GH, Rennie D (eds) Users' guide to the medical literature. American Medical Association, Chicago, pp 449–460

Sackett DL (1969) Clinical epidemiology. Am J Epidemiol 89:125–128

Sackett D (1986) Rules of evidence and clinical recommendations on the use of antithrombotic agents. Arch Intern Med 146:464–465

Sackett D (2002) Clinical Epidemiology: what, who, and whither. J Clin Epidemiol 55:1161–1166

Sackett DL, Winkelstein W (1967) The relationship between cigarette usage and aortic atherosclerosis. Am J Epidemiol 86:264–270

Sackett D, Chambers L, MacPherson A, Goldsmith C, McAuley R (1977) The development and application of indices of health: general methods and a summary of results. Am J Public Health 67:423–428

Sackett DL, Haynes RB, Guyatt GH, Tugwell P (1991) Clinical epidemiology: a basic science for clinical medicine. Little, Brown, Boston

Sackett D, Straus S, Richardson W, Rosenberg W, Haynes RB (2000) Evidence-based medicine. Churchill Livingston, Toronto, p 166

Santesso N, Schunemann H, Blumenthal P, De Vuyst H, Gage J, Garcia F, Jeronimo J, Lu R, Luciani S, Quek SC, Awad T, Broutet N (2012) World Health Organization guidelines: use of cryotherapy for cervical intraepithelial neoplasia. Int J Gynaecol Obstet 118(2):97–102

Schunemann HJ (2011) Guidelines 2.0: do no net harm-the future of practice guideline development in asthma and other diseases. Curr Allergy Asthma Rep 11(3):261–268

Schünemann H, Munger H, Brower S, O'Donnell M, Crowther M, Cook D, Guyatt G (2004) Methodology for guideline development for the seventh American College of Chest Physicians conference on antithrombotic and thrombolytic therapy. Chest 26(3 Suppl):174S–178S

Schunemann HJ, Fretheim A, Oxman AD (2006a) Improving the use of research evidence in guideline development: 1. Guidelines for guidelines. Health Res Policy Syst 4:13

Schunemann HJ, Fretheim A, Oxman AD (2006b) Improving the use of research evidence in guideline development: 10. Integrating values and consumer involvement. Health Res Policy Syst 4:22

Schunemann HJ, Hill SR, Kakad M, Bellamy R, Uyeki TM, Hayden FG, Yazdanpanah Y, Beigel J, Chotpitayasunondh T, Del Mar C, Farrar J, Tran TH, Ozbay B, Sugaya N, Fukuda K, Shindo N, Stockman L, Vist GE, Croisier A, Nagjdaliyev A, Roth C, Thomson G, Zucker H, Oxman AD (2007a) WHO rapid advice guidelines for pharmacological management of sporadic human infection with avian influenza A (H5N1) virus. Lancet Infect Dis 7(1):21–31

Schunemann HJ, Hill SR, Kakad M, Vist GE, Bellamy R, Stockman L, Wisloff TF, Del Mar C, Hayden F, Uyeki TM, Farrar J, Yazdanpanah Y, Zucker H, Beigel J, Chotpitayasunondh T, Hien TT, Ozbay B, Sugaya N, Oxman AD (2007b) Transparent development of the WHO rapid advice guidelines. PLoS Med 4(5):e119

Schunemann HJ, Osborne M, Moss J, Manthous C, Wagner G, Sicilian L, Ohar J, McDermott S, Lucas L, Jaeschke R (2009) An official American Thoracic Society Policy statement: managing conflict of interest in professional societies. Am J Respir Crit Care Med 180(6):564–580

Schünemann HJ, Oxman A, Higgins JPT, Vist GE, Glasziou P, Guyatt GH (2011a) Chapter 11: Presenting results and 'Summary of findings' tables. In: Higgins JPT, Green S (eds) Cochrane handbook for systematic reviews of interventions version 5.1.0 (updated March 2011). The Cochrane Collaboration. Available from http://www.cochrane-handbook.org, 2011

Schünemann HJ, Oxman A, Vist GE, Higgins JPT, Deeks JJ, Glasziou P, Guyatt GH (2011b). Chapter 12: Interpreting results and drawing conclusions. In: Higgins JPT, Green S (eds) Cochrane handbook for systematic reviews of interventions version 5.1.0 (updated March 2011). The Cochrane Collaboration. Available from http://www.cochrane-handbook.org, 2011

Schünemann H, Hill S, Guyatt G, Akl EA, Ahmed F (2011c) The GRADE approach and Bradford Hill's criteria for causation. J Epidemiol Community Health 65:392–395

Shekelle PG, Woolf SH, Eccles M, Grimshaw J (1999) Clinical guidelines: developing guidelines. BMJ 318(7183):593–596

Shekelle P, Woolf S, Grimshaw JM, Schunemann HJ, Eccles MP (2012) Developing clinical practice guidelines: reviewing, reporting, and publishing guidelines; updating guidelines; and the emerging issues of enhancing guideline implementability and accounting for comorbid conditions in guideline development. Implement Sci 7(1):62

Shiffman RN, Shekelle P, Overhage JM, Slutsky J, Grimshaw J, Deshpande AM (2003) Standardized reporting of clinical practice guidelines: a proposal from the Conference on Guideline Standardization. Ann Intern Med 139(6):493–498

Smith D, Baker G, Davies G, Dewey M, Chadwick D (1993) Outcomes of add-on treatment with Lamotrigine in partial epilepsy. Epilepsia 34:312–322

Smith R, Cokkinides V, Eyre H (2003) American Cancer Society guidelines for the early detection of cancer, 2003. CA Cancer J Clin 53:27–43

Soteriades ES, Evans JC, Larson MG, Chen MH, Chen L, Benjamin EJ, Levy D (2002) Incidence and Prognosis of Syncope. N Engl J Med 347(12):878–885

Sox HC, Blatt MA, Higgins MC, Marton KL (1988) Medical decision making. Butterworths, Boston

Stacey D, Bennett CL, Barry MJ, Col NF, Eden KB, Holmes-Rovner M, Llewellyn-Thomas H, Lyddiatt A, Legare F, Thomson R (2011) Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev Oct 5; (10):CD001431

Stacey D, O'Connor AM, DeGrasse C, Verma S (2003) Development and evaluation of a breast cancer prevention decision aid for higher-risk women. Health Expect 6(1):3–18

Stewart M (1995) Effective physician-patient communication and health outcomes: a review. Can Med Assoc J 152:1423–1433

Stiggelbout A, Kiebert G (1997) A role for the sick role. Patient preferences regarding information and participation in clinical decision-making. Can Med Assoc J 157:383–389

Strull W, Lo B, Charles G (1984) Do patients want to participate in medical decision making? JAMA 252:2990–2994

Symons MJ, Moore DT (2002) Hazard rate ratio and prospective epidemiological studies. J Clin Epidemiol 55(9):893–899

Szabo E, Moody H, Hamilton T, Ang C, Kovithavongs C, Kjellstrand C (1997) Choice of treatment improves quality of life. A study on patients undergoing dialysis. Arch Intern Med 157: 1352–1356

Tandon P, Stander H, Schwarz RP Jr (1989) Analysis of quality of life data from a randomized, Placebo controlled heart-failure trial. J Clin Epidemiol 42:955–962

Tarlov A, Ware J, Greenfield S, Nelson E, Perrin E, Zubkoff M (1989) The medical outcomes study. JAMA 262:925–930

The CAPRICORN Investigators (2001) Effects of carvedilol on outcome after myocardial infarction in patients with left ventricular dysfunction: the CAPRICORN randomised trial. Lancet 357:1385–1390

Thomson R, Parkin D, Eccles M, Sudlow M, Robinson A (2000) Decision analysis and guidelines for anticoagulant therapy to prevent stroke in patients with atrial fibrillation. Lancet 355: 956–962

Torrance G (1986) Measurement of health state utilities for economic appraisal. J Health Econ 5:1–30

Tugwell P, Bombardier C, Buchanan W, Goldsmith G, Grace E, Bennett K, Williams J, Egger M, Alarcon GS, Guttadauria M (1990) Methotrexate in rheumatoid arthritis. Impact on quality of life assessed by traditional standard-item and individualized patient preference health status questionnaires. Arch Intern Med 150:59–62

U.S. Department of Health and Human Services (2013) National Guidelines Clearinghouse. http://www.guidelines.gov. Accessed 5 May 2013

USPSTF (2002) Screening for prostate cancer: recommendations and rationale. Ann Intern Med 137:915–916

van Walraven C, Hart R, Singer D (2002) Oral anticoagulants vs aspirin in nonvalvular atrial fibrillation: an individual patient meta-analysis. JAMA 288:2441–2448

Vandvik PO, Lincoff AM, Gore JM, Gutterman DD, Sonnenberg FA, Alonso-Coello P, Akl EA, Lansberg MG, Guyatt GH, Spencer FA (2012) Primary and secondary prevention of cardiovascular disease: antithrombotic therapy and prevention of thrombosis, 9th ed: American College of Chest Physicians Evidence-based clinical practice Guidelines. Chest 141(2 Suppl): e637S–e668S

Ware J, Sherbourne C (1992) The MOS 36-item short-form health survey (SF-36). Med Care 30:473–483

Ware JE Jr, Kozinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A (1995) Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results of the medical outcomes study. Med Care 33:AS264–AS279

Weil J, Colin-Jones D, Langman M, Lawson D, Logan R, Murphy M, Rawlins M, Vessey M, Wainwright P (1995) Prophylactic aspirin and risk of peptic ulcer bleeding. BMJ 310:827–830

Weiss N (1996) Clinical epidemiology: The study of the outcome of illness, 2nd edition. Oxford University Press, Oxford

Whelan T, Sawka C, Levine M, Gafni A, Reyno L, Willan A, Julian J, Dent S, Abu-Zahra H, Chouinard E, Tozer R, Pritchard K, Bodendorfer I (2003) Helping patients make informed choices: a randomized trial of a decision aid for adjuvant chemotherapy in lymph node-negative breast cancer [comment]. J Natl Cancer Inst 95(8):581–587

Wiebe S, Guyatt G, Weaver B, Matijevic S, Sidwell C (2003) Comparative responsiveness of generic and specific quality-of-life instruments. J Clin Epidemiol 56(1):52–60

Woloshin S, Schwartz LM (2011) Communicating data about the benefits and harms of treatment: a randomized trial. Ann Intern Med 155(2):87–96

Woolf S (1994) An organized analytic framework for practice guideline development: using the analytic logic as a guide for reviewing evidence, developing recommendations, and explaining the rationale. In: McCormick KA, Moore S, Siegel RA (eds) Methodology perspectives. US Department of Health and Human Services, Agency for Health Care Policy and Research, Washington, DC, pp 105–113

Woolf SH, Grol R, Hutchinson A, Eccles M, Grimshaw J (1999) Clinical guidelines: potential benefits, limitations, and harms of clinical guidelines. BMJ 318(7182):527–530

World Health Organization (2003) Global programme on evidence for health policy. Guidelines for WHO guidelines. EIP/GPE/EQC/2003.1. World Health Organization, Geneva

Pharmacoepidemiology

48

Edeltraut Garbe and Samy Suissa

Contents

E. Garbe (✉)
Department of Clinical Epidemiology, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

Faculty of Human and Health Sciences, University of Bremen, Bremen, Germany

S. Suissa
Division of Clinical Epidemiology, McGill University Health Center, Montreal, QC, Canada

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_31,

48.1 Introduction

In the last decades we have witnessed a tremendous progress in the medical sciences that has led to the development of a great number of new powerful pharmaceuticals. These new medicines enable us to provide much better medical care, but occasionally they will cause harm and give rise to serious adverse reactions that were unexpected from preclinical studies or premarketing clinical trials. Against this background, pharmacoepidemiology has developed as a scientific discipline at the interface between clinical pharmacology and clinical epidemiology (cf. chapter ►Clinical Epidemiology and Evidence-Based Health Care of this handbook). Pharmacoepidemiology can be defined as the application of epidemiological knowledge, methods, and reasoning to the study of the effects and uses of drugs in human populations (Porta et al. 1997). The application of epidemiological methods – i.e., the use of non-experimental observational techniques – the epidemiological perspective with an emphasis on investigations in large unselected populations and long-term studies, the public health approach, and the philosophy of epidemiology are all extended to the scope of clinical pharmacology, i.e., the study of the effects of pharmaceuticals in humans.

Pharmacoepidemiology investigates both beneficial and adverse drug effects. Its focus and the one that receives the greatest attention is the assessment of the risk of uncommon, at times latent, and often unexpected adverse reactions that present for the first time after a drug has been marketed. The greatest challenge of pharmacoepidemiology is then to quantify the risk of a drug accurately, relative to one or several alternatives.

The study of adverse drug effects poses a number of methodological difficulties that must be addressed by pharmacoepidemiological research designs: First, drug exposure is not a stable phenomenon. Drug prescription habits may change due to the development of new pharmaceuticals, better knowledge on already available medications, or other reasons. Second, drug exposure can be sensitive to a great number of factors that may also be related to the outcome of interest, such as the indication for prescribing potential contraindications for drug use, the natural course of the disease or disease severity and compliance. Third, the risk of an adverse drug reaction (ADR) is often not constant, but it may change over time which may have important implications for the design and interpretation of pharmacoepidemiology studies.

In this chapter, we will first discuss limitations of premarketing clinical trials; we will then describe the characteristics of spontaneous reporting systems which have been implemented by regulatory agencies and the pharmaceutical industry for postmarketing surveillance. Another issue will be the use of multipurpose cohorts and large administrative health-care databases for drug effect studies, which have found widespread application in pharmacoepidemiology. We will further discuss several methodological aspects that are unique to pharmacoepidemiological research, e.g., the phenomenon of “depletion of susceptibles,” which is a form of selection bias, or “confounding by indication,” which is also referred to as “confounding by disease severity” or “channelling bias.” Unmeasured confounding is of particular concern in pharmacoepidemiology studies conducted with large healthcare utilization databases which often do not contain information on all potential confounders. We will discuss instrumental variable analyses as an approach to study drug effects in the situation of unmeasured confounders, if a valid instrument can be found. We will further present the use of propensity scores, as a tool to reduce confounding particularly in studies of intended drug effects, and will discuss newer approaches of studying drug effects, such as the case-crossover, case-time-control, and self-controlled case series designs. Characteristics of drug utilization studies and their units of measurement will also be discussed.

48.2 Limitations of Premarketing Clinical Trials

Prior to marketing, new drugs are subjected to preclinical animal studies followed by three phases of clinical trials in humans. These phases are divisions of convenience in what is a continuous process of acquiring knowledge on the effects of a new drug in humans. In *phase I studies*, humans are exposed to a new drug for the first time. These studies are often conducted in small numbers of healthy volunteers and are intended to explore the tolerability, pharmacokinetic, and pharmacodynamic properties of a new drug in humans. In *phase II studies*, the optimal dose range of the new drug is investigated, and its efficacy and safety are explored in the intended patient population. These studies usually include several hundreds of patients. *Phase III studies* are aimed to prove the efficacy and safety of the new drug under strictly controlled experimental conditions in a larger patient population. They are mostly conducted as randomized controlled clinical trials (RCT) and often include several thousand patients altogether. In spite of the size of phase III studies, these studies have still limited ability to identify rare ADRs, since this would require an even larger number of study participants.

Table 48.1 displays sample size calculations for prospective studies. It shows the number of patients needed to detect a relative risk of a given magnitude in relation to the incidence of the event in the reference group. We can see that for the detection

Table 48.1 Sample sizes for detection of a given drug risk in an RCT or a cohort study (size per study arm)

	Incidence of outcome in the control group			
Relative risk	1/50,000	1/10,000	1/5,000	1/1,000
2.0	1,177,295	235,430	117,697	23,511
2.5	610,446	122,072	61,025	12,187
3	392,427	78,472	39,228	7,832
5	147,157	29,424	14,707	2,934
7.5	78,946	15,783	7,888	1,572
10	53,288	10,652	5,323	1,059

Calculations are based upon a two-sided significance level α of 0.05, a power of 80% ($\beta = 0.2$), and one control subject per exposed subject

Table 48.2 Drug removals from the US market between 1997 and 1998. Number of patients exposed to withdrawn drugs in clinical trials compared to actual use after marketing

Removed drug	Number of patients exposed before marketing[a]	Approximate exposure prior to withdrawals
Terfenadine	5,000	7,500,000
Fenfluramine	340	6,900,000
Dexfenfluramine	1,200	2,300,000
Mibefradil	3,400	600,000
Bromfenac	2,400	2,500,000

[a] Number of patients included in the US premarketing studies

of very rare ADRs with sufficient statistical power, prohibitively large sample sizes in premarketing clinical studies would be needed. It is thus inherent in the drug development process – taking into account the already high cost for development of new pharmaceuticals – that serious rare ADRs will usually only be detected after drug marketing when the drug has been used in large patient populations.

An investigation by the Food and Drug Administration (FDA) into the withdrawal of five pharmaceuticals from the US market between 1997 and 1998 illustrates this point (Friedman et al. 1999). All five drugs were removed from the US market because of the discovery of unexpected serious adverse drug reactions (ADR) in the postmarketing period. The FDA investigated whether this unexpectedly high number of drug removals in only a 12-month period was related to the expedited drug review and approval process that had been implemented. They calculated the number of patients exposed in the clinical trials before marketing and the approximate number of patients exposed before drug withdrawal (Table 48.2). The figures demonstrate that usually huge numbers of patients need to be exposed before sufficient knowledge on a rare ADR has been accumulated. For example, serious hepatotoxic effects of bromfenac occurred in approximately 1 in 20,000 patients who took the drug for longer than 10 days (Friedman et al. 1999). To reliably detect this toxic effect, some 100,000 patients would need to be included in the premarketing clinical studies.

In addition, premarketing clinical studies differ from routine clinical care for a number of other reasons: (1) These studies mostly include a selected study population, defined by strict inclusion and exclusion criteria, which is often not fully representative of subsequent users of the drug. It is well known that premarketing studies tend to under represent the elderly, patients with comorbid conditions, pregnant women, and children. Patients in premarketing trials may even be considered a selected group of patients just because they are willing or able to participate. (2) Premarketing clinical trials are performed at selected sites which are typically better equipped than routine care facilities. They are conducted by specialists in their field, and all participating persons have been specially informed and trained. Surveillance of patients is almost by definition more intensive than in routine clinical treatment, if only one considers the frequency and spectrum of laboratory tests or the assessment of therapeutic and unwanted effects. (3) Treatment regimens in premarketing clinical trials are largely fixed and allow almost no individual treatment variations. In contrast, adjustments are constantly made in routine care, depending on the progress of therapy and on the interaction between doctor and patient. (4) Premarketing clinical trials are usually of short duration. This renders it impossible to detect ADRs that only develop after a long induction period or after cumulative drug intake.

For all these reasons, crucial answers to questions of drug safety cannot be provided even by the most valid and complex phase III study.

48.3 Characteristics of Spontaneous Reporting Systems

48.3.1 Description

In the early 1960s, systems evolved in most Western countries that collected spontaneous reports on ADRs from doctors. Establishment of these spontaneous reporting systems was largely a consequence of the "thalidomide disaster," in which children exposed to the hypnotic thalidomide in utero were born with phocomelia, a congenital deformity of the limbs resulting from prenatal interference of the drug with the development of the fetal limbs (Wiholm et al. 2000). Worldwide, several thousand cases of limb malformations in newborns observed in the 1950s and 1960s were attributed to the use of thalidomide during pregnancy (Lenz 1987). Based on this experience, spontaneous reporting systems were set up to monitor drug safety in the postmarketing period. In these systems, physicians – in some countries also pharmacists, other health-care professionals, or patients – report the suspicion of an ADR to the country's drug regulatory agency or to the pharmaceutical company that is marketing the drug. Drug regulatory agencies exchange the ADR reports with the concerned pharmaceutical companies and vice versa. The reports are locally assessed; the reported adverse event terms are coded using a standardized international terminology, e.g., MedDRA (the Medical Dictionary for Regulatory Activities), and are entered in a computerized database. In December 2011, 103 countries also forwarded their ADR reports to the World Health Organization (WHO) Collaborating Centre for International Drug Monitoring in Uppsala (Uppsala Monitoring Centre 2011). This so-called Vigibase^{TM} set up by the WHO contained already over seven million ADR reports at that time. The ADR reports are subject to further statistical screening to identify possible new risks associated with drugs and are further analyzed by a review team. When there are a number of reports of a suspected new adverse reaction to a particular drug, this may be circulated among program members as a "signal" – a notice of a need for increased awareness of a possible hazard. Through membership in the WHO program, one country can know whether similar ADR reports are being made elsewhere. An ADR report usually contains the patient's demographic information including age and sex; the patient's weight and height; adverse event (AE) information including date and outcome of the event, description of the event, evidence, and existing medical history; information of suspect and concomitant medicine(s), including drug name, dose, application, and indication for use; whether the event abated after drug use was stopped ("dechallenge") and whether the event reappeared after the drug was reintroduced ("rechallenge"); and the reporter's name and address.

Each report is assessed for its causality with drug intake by trained reviewers. Causality assessment is usually based on a set of criteria which includes the time interval between drug administration and the onset of the ADR, the course of the reaction when the drug was stopped, the results of re-administration of the drug, the existence of other causes that could also account for the observed reaction such as patient comorbidity or concomitant drug treatment, the pattern of the adverse effect, and the existence of reliable and specific laboratory test results. Causality assessment is not based on the single case report, but will also take into account other available information on the drug(s). In France, causality assessment criteria have been built into an algorithm (the "official method of causality assessment") that is used throughout the country (Benichou 1994). The French method distinguishes "intrinsic imputability" which takes into account only the single case report information from "extrinsic imputability" which is based on all published data on all drugs. Overall, causality assessment from individual case reports is a complex task and often associated with a high degree of uncertainty, since confounding by concomitant drug therapy or the underlying disease can frequently not be ruled out. Rare exceptions are a positive rechallenge to the drug (which is mostly accidental and involuntary, since this is rarely without risks) or positive results of specific laboratory tests such as the detection of drug-dependent antibodies.

Spontaneous reporting systems have several important advantages. They are relatively inexpensive to operate with respect to staff and basic technical equipment. They have the potential to cover the whole patient population and are not restricted to either hospitalized patients or outpatients. Surveillance starts as soon as a drug is marketed, and monitoring of drug safety continues throughout the whole post-marketing life cycle of a drug. The suspicion of an ADR is, in theory, based on the experience of all treating physicians and pharmacists. Spontaneous ADR reporting systems can provide an alert to very rare, but nevertheless potentially important drug toxicity. Spontaneous reporting systems have identified many new, i.e., previously unrecognized drug hazards, e.g., clozapine-induced granulocytopenia, captopril-induced cough, and amiodarone-induced hepatotoxicity. Further examples can be found in the article by Rawlins et al. (1989).

48.3.2 Limitations

Many of the successes of spontaneous reporting systems have been in the recognition of ADRs occurring shortly after starting therapy. Spontaneous reporting schemes are much less effective in identifying reactions with a long induction period. An example is the oculomucocutaneous syndrome associated with exposure to practolol which was undetected by the yellow card spontaneous reporting scheme in the UK (Venning 1983). For the same reason, spontaneous reporting schemes are not well suited to identify drug-induced cancerogenicity.

ADR reports often do not provide sufficient information to confirm that a drug caused an event. For example, the ADR report may not give enough details on comorbidity or other medications to rule out other possible causes for the event in a remote expert assessment. It may be impossible to exclude confounding by indication, i.e., that the cause for the reported adverse event is rather related to the indication for the drug than to the drug itself. As an example, depression has been reported with the anti-acne medication roaccutane (Wysowski et al. 2001). Depression could, however, also be related to psychological disturbances over severe acne in sensitive teenagers rather than to roaccutane itself.

Recognition of an ADR depends on the level of diagnostic suspicion of the treating physician and may be related to the nature of the adverse event. Some ADRs are more likely to be diagnosed and reported than others because of their known association with drug therapy. For example, acute agranulocytosis is attributable to drug treatment in about 60–70% of cases (Kaufman et al. 1996). It may therefore be more likely to be attributed to drug therapy than a disorder, e.g., acute myocardial infarction, that is usually not related to drug treatment (Faich 1986).

Spontaneous reporting systems suffer from serious underreporting of ADRs and various biases that affect reporting. Even in the UK, a country with a relatively high reporting rate in relation to its population size, rarely more than 10% of serious ADRs are notified to the regulatory agency (Rawlins et al. 1989). In France, a recent comparison of ADR reports with data about drug-induced hospitalizations in three pharmacoepidemiology field studies indicates that only 5% of ADRs leading to hospitalizations are actually reported in the spontaneous reporting system (Begaud et al. 2002). Lack of knowledge how to report an ADR and misconceptions about the type of ADR that should be reported are important reasons contributing to underreporting (Eland et al. 1999). But it also has to be taken into account that spontaneous reporting and published case reports may also lead to numerous false alarms.

Medical or mass media attention can stimulate reporting in a distorted manner and give rise to differential reporting in a dramatic way (Griffin 1986). An example of such “media bias” is that of central nervous side effects following treatment with the benzodiazepine triazolam. After van der Kroef published a case series of these ADRs in 1979 (van der Kroef 1979), these side effects received extensive media coverage on Dutch television. As a consequence, the Netherlands received 999 ADR reports related to triazolam in 1979 out of a total of 1912 annual reports in 1979 overall (Griffin 1986). Reporting can also be affected by the market share of the drug, the quality of the manufacturer’s surveillance system, reporting regulations, and the length of time a drug is on the market (Griffin 1986; Lindquist and Edwards 1993). It has been shown that reporting rates do not remain stable over time, but usually rise until approximately the middle to end of the second year of marketing and then progressively decline over the following years despite steadily increasing prescribing rates. This phenomenon was first described by Weber in the context of a study of nine non-steroidal anti-inflammatory drugs (NSAIDs) in the UK during the late 1970s and early 1980s and has therefore been called the "Weber effect" (Weber 1984). The Weber effect has also been reported from other studies of adverse drug reactions (Haramburu et al. 1992, 1997; Hartnell and Wilson 2004).

48.3.3 Statistical Approaches: Reporting Rates and Disproportionality Measures

An ideal early warning system should not only recognize new hazards but also provide an estimate of their incidence. Spontaneous reporting systems may provide alerts of drug hazards, but they cannot be used to calculate incidence rates of adverse events related to a specific drug. Calculation of an incidence rate requires accurate numerator and denominator information, both of which are not available from the spontaneous reporting systems. First, the extent of underreporting for an individual drug is very difficult to assess and may even differ between drugs of the same pharmacological class. Second, the population exposed to the drug (i.e., the population at risk) is unknown and cannot be determined from drug sales data, since the duration of drug use, the dose regimen, and compliance in individual patients are unknown.

Instead of the calculation of incidence rates, reporting rates (number of AE reports per market share) based on sales data are sometimes computed as an alternative approach (Moore et al. 2003; Pierfitte et al. 2000). Calculation of reporting rates is based on the assumption that the magnitude of underreporting is reasonably similar for similar drugs that share the same indication, country, and period of marketing (Pierfitte et al. 1999). The comparison of ADR reporting rates should therefore be restricted to drugs of the same category used for the same indication. Factors that may bias the comparison of reporting rates include differences in the length of time the drugs are on the market, differences in exposure populations, secular reporting trends, reporting variations, diagnosis and prescription variation, and the publicity of an ADR. Statistical corrections for year of marketing, secular trends of all-drug-all-adverse event reporting, and drug usage have been proposed (Tsong 1995). These adjustments do not, however, cover all possible sources of bias. In particular, do they not erase concerns about differences in the magnitude of underreporting for different drugs: Interpretation of reporting rates should therefore be conducted with an understanding of their limitations. Differences in reporting rates do not establish differences in incidence rates. They may, however, provide alerts of drug hazards to be investigated by more rigorous pharmacoepidemiological study designs.

The proportional reporting ratio (*PRR*) and the reporting odds ratio (*ROR*) have been proposed as measures of disproportionality for signal generation (Evans et al. 2001; Stricker and Tijssen 1992). These measures are based on the observed number of reports of a particular suspected drug-ADR combination compared with an estimate of the expected number based on all other reporting in the same database; the expected number of reports is calculated on the assumption of independence of the drug and the suspected ADR. The ADR databases maintained by the regulatory authorities and the WHO contain a large number of reports suitable for statistical signal generation, e.g., over 7 million reports in the WHO database (Uppsala Monitoring Centre 2011), over 4 million reports in the Adverse Event Reporting System (AERS) database maintained by the FDA (US Food and Drug Administration 2010), and over 600,000 reports in the Yellow Card Scheme of the Medicines and Healthcare Products Regulatory Agency (MHRA) (MHRA Centre NorthWest 2009). The *PRR* or *ROR* involves calculation of the proportions of specified reactions or groups of reactions for drugs of interest where the comparator is all other drugs or other selected drug groups in the database. Whereas the *PRR* is the quotient of $a/(a+c)$ divided by $b/(b+d)$ derived from the reported frequencies of all drug-event pairs in the database arranged in a two-by-two table, the *ROR* is the quotient of a/c divided by b/d (Table 48.3). In practice both measures tend to be similar, since usually $a << (b, c) << d$ (Almenoff et al. 2007).

The *PRR* and *ROR* behave in a similar fashion as the relative risk, i.e., the higher the *PRR* or *ROR*, the greater the strength of the signal. The usual application of these measures is in conjunction with computing the two-by-two table chi-square values for association on 1 degree of freedom. With this approach, large ratios of the disproportionality measures with non-significant chi-square values, which are typically seen with small numbers of suspected drug–ADR combinations and even smaller expected values, are ignored (Almenoff et al. 2007). Signals can then be identified based on the value of the *PRR* or *ROR*, the value of the chi-squared test, and the absolute number of reports. In a proof-of-concept study, Evans et al. (2001) defined a signal as a *PRR* value of at least 2, a value of chi-squared test of at least 4, and a minimum of 3 cases. Using these criteria on the UK Yellow Card Scheme database for 15 newly marketed drugs, they identified 481 potential signals, 339 (70%) of which were recognized ADRs, 62 (13%) were considered to be related to the underlying disease, and 80 (17%) were signals requiring further evaluation. Statistical approaches such as calculation of *PRR*s or *ROR*s are not a substitute for a detailed ADR review, but they may aid in the decision on which series of cases should be investigated next. Similarly, as already mentioned for reporting rates, the *PRR* or *ROR* may be affected by differential ADR reporting related to notoriety, surveillance, and market size effects. A *PRR* or *ROR* above 1 may therefore just indicate a higher reporting of a possible reaction under a drug, but not necessarily a differential occurrence (Moore et al. 2003). Whereas for calculation of the *ROR* cells a, b, c, and d has to contain reports, only reports in cells a and b are required for calculation of the *PRR* (van Puijenbroek et al. 2002). With the *ROR*, different adjustments can be calculated in logistic regression analysis, and interaction terms can be included in the model (van Puijenbroek et al. 2002). A discussion of both measures can be found in Rothman et al. (2004) and Waller et al. (2004).

Table 48.3 Calculation of the proportional reporting ratio (*PRR*) or the reporting odds ratio (*ROR*) (a, b, c, and d are absolute frequencies of the according combination)

	Drug of interest	All other drugs in the database
Event of interest	a	b
All other events	c	d

Other measures of disproportionality have been proposed using Bayesian statistics (Bate et al. 1998; DuMouchel 1999; Szarfman et al. 2002). These more complex statistical approaches are, like the *PRR* or *ROR*, based on a comparison of observed versus expected frequencies of ADRs under a particular drug, but they differ in the way they relate all drug-event combinations in the database to each other. Bayesian approaches account for the uncertainty in the disproportionality ratio when the counts are small. Bayesian statistical methods produce "shrinkage" values of the observed/expected ratio, so that the raw value of the observed/expected ratio is moved toward the null hypothesis value of 1 by an amount that depends on the variability of the disproportionality statistic (DuMouchel 1999). The raw values of the observed/expected ratio are transformed using Bayesian theory to a basic estimate called *EBGM* (Empirical Bayes Geometric Mean of the relative reporting ratio), and a 90% confidence interval (*EB05*, *EB95*) is calculated. In the WHO database, a similar procedure is used which calculates a statistic known as the Information Component (Bate et al. 1998, 2008).

As more data become available, the shrinkage property will gradually diminish in influence. Shrinkage is often negligible when there is a large observed count for a drug-event combination. Shrinkage thus reduces the number of false-positive associations which would be generated for combinations with little observed data where the observed/expected ratio might be unexpectedly high for the very low expected counts of rarely reported drugs and rarely reported ADR terms (Bate et al. 2008).

48.4 Sources of Data in Pharmacoepidemiological Research

A great number of pharmacoepidemiology studies are being conducted as field studies, with data being collected for the specific hypothesis under study. These studies are sometimes conducted in an international setting to increase the number of cases and to provide more timely results (Abenhaim et al. 1996; Anonymous 1986, 1995a; Spitzer et al. 1996). Increasingly, already existing data sources are being used for pharmacoepidemiological research. Existing data sources include multipurpose cohort studies or large health-care utilization databases. Studies utilizing such data can be conducted more quickly and are less expensive than field studies, since the data have already been collected.

48.4.1 Multipurpose Cohorts

Multipurpose cohorts are designed to investigate many different research hypotheses. Their study population usually consists of a subset of a defined population that has not been assembled by a specific exposure, but by other factors. For example, in the US Nurses' Health Studies, study participants were defined by age, female sex, and profession (Nurses' Health Study I: 121,700 female nurses aged 30–55 years at baseline in 1976; Nurses' Health Study II: 116,671 female nurses aged 25–42 years at baseline in 1989). If a multipurpose cohort is used to investigate an association between a specific drug exposure and a disease, its cohort members will usually have sufficient variability in their exposure status for the drug to be investigated: They may currently be exposed or unexposed, they may be exposed to different doses of the drug, they may have been exposed in the past, or they may be exposed in the future. If, in addition, disease occurrence and relevant confounder information has been ascertained, the multipurpose cohort data may be used to investigate a specific pharmacoepidemiological hypothesis. The US Nurses' Health Studies have been extensively used for pharmacoepidemiology research questions. Examples include the association between nonsteroidal anti-inflammatory drugs and risk of Parkinson's disease (Chen et al. 2003); use of estrogens and progestins and risk of breast cancer in postmenopausal women (Colditz et al. 1995); postmenopausal estrogen and progestin use and risk of cardiovascular disease (Grodstein et al. 1996); oral contraceptives and the risk of multiple sclerosis (Hernan et al. 2000); aspirin, other non-steroidal drugs, and risk of ovarian cancer (Fairfield et al. 2002); calcium intake and risk of colon cancer (Wu et al. 2002); and many more associations (Grodstein et al. 1998; Hee and Grodstein 2003; Hernandez-Avila et al. 1990; Weintraub et al. 2002). Other multipurpose cohorts that have been less frequently used for pharmacoepidemiological research include the Health Professionals Follow-Up Study (Chen et al. 2003; Giovannucci et al. 1994; Wu et al. 2002), the National Health and Nutrition Examination Survey I (NHANES) epidemiological follow-up study (Funkhouser and Sharp 1995; Lando et al. 1999), the Framingham cohort study (Abascal et al. 1998; Felson et al. 1991; Kiel et al. 1987; Worzala et al. 2001), and the Rotterdam Study (Beiderbeck-Noll et al. 2003; Feenstra et al. 2002; Schoofs et al. 2003).

48.4.2 Record Linkage Studies

Large health-care utilization databases have emerged as another important data source for pharmacoepidemiology research. In the United States and Canada, administrative health-care utilization databases have been set up for the administration of reimbursement payments to health-care providers in nationally funded health-care systems or managed care organizations. In the Scandinavian countries, data are based on reimbursement data in state-funded health-care systems. These databases offer the advantage that their data can be linked to birth, cancer, and other disease registries as well as to mortality registries based on the pseudonymized unique personal identifier used in these countries (Furu 2008; Furu et al. 2010). In the United Kingdom, the Netherlands (IPCI 2012a), and some other countries, large health-care utilization databases consist of data entered by general practitioners (GP) into their practice computers (Schneeweiss and Avorn 2005; Suissa and Garbe 2007). In Germany, a large population-based database has recently been established based on data from several health insurances (Pigeot and Ahrens 2008) and has meanwhile been applied in several pharmacoepidemiological studies (Amann et al. 2012; Behr et al. 2010; Garbe et al. 2011; Jobski et al. 2011). A brief overview of selected databases is given in Table 48.4. More detailed information on some of these databases can be found in the textbook "Pharmacoepidemiology" by Strom et al. (2012).

Table 48.4 Examples of administrative health databases in the USA, Canada, and Europe

Database	Characteristics	Eligible population	Drug prescriptions and/or dispensations since
Saskatchewan's Health Databases, Saskatchewan, Canada (Saskatchewan Ministry of Health 2010)	Provincial health plan	1 million	1975
RAMQ database, Quebec, Canada (RAMQ 2010)	Provincial health plan	7.6 million[a]	1997
Group Health Cooperative, Washington, USA[b] (Group Health 2012)	Health Maintenance Organizations (HMO)	600,000	1977
Kaiser Permanente, Northern California, USA (Kaiser Permanente 2012a)	HMO	3.2 million	1994 (from all pharmacies)
Kaiser Permanente Northwest Division, USA (Kaiser Permanente 2012b)	HMO	479,000	1986
Tennessee Medicaid database, USA[c] (Tennessee Government 2012)	Health insurance for low-income children, parents, pregnant women, and elderly and disabled adults	1.2 million	1977
New Jersey Medicaid Database, USA (State of New Jersey 2012)	Health insurance for low-income children, parents, pregnant women, and elderly and disabled adults	1.2 million	1980
CPRD (former GPRD), UK (GPRD 2012)	Medical records from primary care	13 million (5.2 million active patient records)	1987 for selective practices, 1991 for the majority
The Health Improvement Network (THIN) database, UK (UCL Research Department of Primary Care and Population Health 2011)	Medical records from primary care	9.1 million (3.4 million active patient records)	1987
QResearch, UK (QResearch 2012b)	Medical records from primary care	13 million (5 million active patient records)	1988
GePaRD (German Pharmacoepidemiological Research Database) (Pigeot and Ahrens 2008)	Reimbursement data from 4 health insurance providers (regional as well as national)	15 million	2004
Integrated Primary Care Information Project (IPCI), Netherlands (IPCI 2012a)	Medical records from primary care	700,000 patients	1994

(continued)

Table 48.4 (continued)

Database	Characteristics	Eligible population	Drug prescriptions and/or dispensations since
PHARMO, Netherlands (PHARMO Institute for Drug Outcomes Research 2012)	Record linkage of pharmacy data with hospital and other data	3.2 million community-dwelling inhabitants of 65 municipal areas in the Netherlands	1998
Pedianet, Italy[d] (Pedianet Project 2012)	Medical records from pediatricians	180,000	2000
Swedish Health and Welfare Statistical Databases (The National Board of Health and Welfare (Socialstyrelsen) 2012a)	Data from the National Health System including prescription data; hospital discharge data; data from birth, cancer, and mortality registries	9.2 million	2005
The National Patient Register (NPR), Sweden (The National Board of Health and Welfare (Socialstyrelsen) 2012b)	Medical records from inpatient care, from 2001 including outpatient care	9.2 million	1987
Norwegian Prescription Database (NorPD) (Furu 2008)	Information on drugs dispensed by prescription from all pharmacies in Norway	4.9 million	2004
STAKES – National Research and Development Centre for Welfare and Health, Finland (Furu et al. 2010)	Statistical and register information on social welfare and health care, diseases, and treatment of diseases	5.3 million	1993
KELA, Finland (KELA – The Social Insurance Institution of Finland 2012)	Data from the Finnish prescription reimbursement register	5.3 million	1994
Danish Registry of Medicinal Product Statistics (The Danish Medicines Agency 2010)	Data from the Danish prescription register	5.5 million	1994
Odense University Pharmacoepidemiological Database (OPED), Denmark (Sturkenboom 2007)	Public institution research registry of dispensing claims data	1.2 million	1990
The Pharmacoepidemiological Prescription Database of North Jutland (PDNJ), Denmark (Sturkenboom 2007)	Public institution research registry of dispensing claims data	1.7 million	1989

[a] Including 3.3 million covered by the Public Prescription Drug Insurance plan

[b] Members from Washington State and Idaho

[c] Tennessee withdrew from the federal Medicaid program on 1/1/94 and implemented TennCare

[d] There was a test period of 3 years. As of August 2007 there were 130 participating family pediatricians

48.4.2.1 Administrative Databases in the USA and Canada

These databases usually consist of patient-level information from two or more separate files which can be linked via a unique patient identifier contained in each file. The unique patient identifier often consists of the social security number of the patient which is "scrambled" to ensure patient confidentiality. Information contained in the different files usually consists of demographic patient information, information on drug dispensations from pharmacies, information on hospitalizations, and information on ambulatory physician visits (Fig. 48.1). Through record linkage, person-based longitudinal files can be created for particular research questions. In some databases, record linkage is possible with cancer registries or birth malformation registries to investigate hypotheses of drug carcinogenicity or teratogenicity. Researchers usually have to submit a study protocol for review by an ethics committee, and they only receive subsets of the files which are extracted to investigate the particular research hypothesis. Fees are charged for the time needed to extract the necessary data from the entire database. All statistical analyses are done on the anonymized data.

Saskatchewan's Health Databases have been used extensively for pharmacoepidemiological studies (Saskatchewan Ministry of Health 2010). These databases will be used to illustrate which information may be expected in an administrative healthcare database (Table 48.5). Health Databases in Saskatchewan are based on the universal health insurance program in this Canadian Province. Differently from the Medicaid program, there is no eligibility distinction based on socioeconomic status. Record linkage is possible with the province's cancer registry. Medical records in hospitals are accessible upon approval from individual district health boards and affiliated facilities. Physician records may also be accessed for specific studies. A wide range of conditions has been validated by hospital chart review, including rheumatoid arthritis (Tennis et al. 1993), hip fractures (Ray et al. 1989), gastrointestinal bleeding (Raiford et al. 1996), and asthma-related conditions (Spitzer et al. 1992).

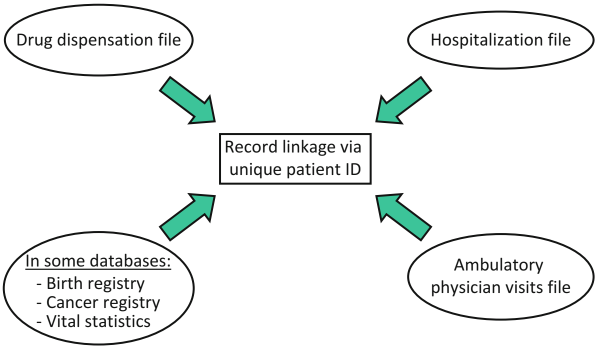


Fig. 48.1 Record linkage in administrative health databases

Table 48.5 Information contained in different files of Health Databases in Saskatchewan (Saskatchewan Ministry of Health 2010)

Population registry	Prescription drug data	Hospital services data	Medical services data
• Name • Health services number (HSN) • Sex • Date of birth • Residence information • Dates of coverage initiation and termination • Reason for coverage termination[g] • Indicator for registered Indian status • Indicator for current social assistance recipients receiving extended health benefits	*Patient information* • HSN • Sex, year of birth • Designation of special status[a] *Drug information* • Pharmacologic-therapeutic classification of drug[b] • Drug identification number[c] • Drug active ingredient number[h] • Generic and brand names • Strength and dosage form • Date dispensed • Quantity dispensed *Provider information* • Prescriber identification number[k] • Dispensing pharmacy identification number *Cost information* • Unit cost of drug materials • Dispensing fee and markup • Consumer share of total cost • Government share of total cost • Total cost	*Patient information* • HSN • Sex, year, and month of birth • Residence *Diagnostic and treatment information* • Most responsible diagnosis • Other diagnoses[d] • Principal procedure • Other procedures[d] • Accident code[i] *Other* • Admission and discharge dates • Length of stay • Admission and separation types • Case mix group • Resource intensity weight • Attending physician • Attending surgeon[e] • Hospital identification number	*Patient information* • HSN • Age, sex • Location of residence • Indicator for registered Indian status *Physician information* • Physician specialty • Referring physician[e] • Physician identification number[f] *Services and diagnostic information* • Date of service • Service code • Type of service • Diagnosis • Location of service[j] • Payment information[l]

[a] e.g., Saskatchewan Assistance Plan recipient

[b] Based on the American Hospital Formulary Service classification system

[c] DIN – assigned by Health Canada

[d] Number potentially available varies depending on the time period

[e] If applicable

[f] Which can be linked with the physician registry for additional information such as physician's age, sex, place and year of graduation, and practice type

[g] e.g., death, left the province

[h] AIN – also assigned by Health Canada

[i] External cause code

[j] e.g., office, inpatient, outpatient, home, other

[k] Which can be linked with the physician registry for additional information such as location of practice and prescriber specialty

[l] e.g., amount paid, date of payment

48.4.2.2 Physician-Based Databases

The former General Practice Research Database (GPRD) is a large physician-based computerized database of anonymized longitudinal patient records from hundreds of general practices in the UK, containing more than 35 million patient years of data. In March 2012, the GPRD became part of the data services provision of the Clinical Practice Research Datalink (CPRD) (Clinical Practice Research Datalink 2012). Currently, information is collected on approximately 3 million patients, equivalent to approximately 5% of the UK population. The database was created in June 1987 as the Value Added Medical Products (VAMP) research databank. VAMP provided practice computers and general practice software to general practitioners (GPs), and in return, GPs consented to undertake data quality training and to contribute anonymized data to a central database for subsequent use in public health research. During the 1990s, VAMP research databank underwent several organizational and management changes. The database was renamed General Practice Research Database (GPRD) in 1994 when it was donated to the UK Department of Health. In 1999, management responsibility for the database was transferred to the UK Medicines Control Agency which became part of the newly created Medicines and Healthcare Products Regulatory Agency (MHRA). The database has been used extensively for pharmacoepidemiology and clinical epidemiology research. A bibliography of studies using GPRD data can be found on the webpage of the CPRD (Clinical Practice Research Datalink 2012). The database includes the following information: demographics, including age and sex of patient (information on race is not collected); medical diagnosis, including comments; all prescriptions; events leading to withdrawal of a drug or treatment; referrals to hospitals; treatment outcomes, including hospital discharge reports where patients are referred to hospital for treatment; and miscellaneous patient information, e.g., smoking status, height, weight, immunizations, and for a growing number of patients also lab results. Recently, the GPRD gained approval to enable record linkage of GPRD data with other UK health-care databases via the patient's NHS (National Health Service) number, sex, date of birth and post code. Specifically, the Hospital Episode Statistics (HES) database records information on all hospitalizations, including data on length of stay, ward types, as well as extensive disease and procedure coding. The linkage between the GPRD and the HES databases applies to approximately half of the practices contributing to the GPRD. Validation studies of the GPRD have shown that the recording of medical data into GPs' computers is almost complete (Garcia Rodriguez and Perez 1998).

Besides the CPRD, other physician-based databases are The Health Improvement Network (THIN) database (The Health Improvement Network 2012) and the QResearch database (QResearch 2012a) in the UK and the IPCI database (IPCI 2012b) in the Netherlands. The MediPlus databases from IMS Health are also physician-based databases which are available in different countries. Depending on the particularities of the respective health-care system, different data are available for research. A description of the German IMS Disease Analyzer–MediPlus database can be found in an article by Dietlein and Schroder-Bernhardi (2002). The IMS databases have not been used extensively for pharmacoepidemiology research. Studies which examine data validity and comprehensiveness are mostly lacking. The German IMS Disease Analyzer–MediPlus database does not contain patient hospitalization data. It also lacks diagnostic or treatment information from all physician specialists, since it is usually based on one panel of doctors only, e.g., on a panel of GPs and internists or gynecologists or urologists. The database derived from the panel of gynecologists would therefore not include information on ambulatory physician contacts in GPs' or internists' offices and vice versa. The German IMS Disease Analyzer–MediPlus database has been used for several drug utilization studies, e.g., on the dosing of cava-cava extracts (Dietlein and Schroder-Bernhardi 2003), whether hospitals influence the prescribing behavior of general practitioners (Schroder-Bernhardi and Dietlein 2002), or how doctors treat Helicobacter pylori infections (Perez et al. 2002). In the UK, the MediPlus database contains similar patient information as the CPRD database.

48.4.3 Advantages and Limitations

Large health-care utilization databases offer several important advantages: (1) They are usually of enormous size, with patient numbers ranging from several hundred thousand to well over several millions. This makes it possible to study rare adverse events of pharmaceuticals in large populations. (2) Medication information is usually more accurate than self-recorded exposure information. Drug histories obtained from patients have limited reliability particularly for drugs that are used only intermittently and not on a regular basis (Kelly et al. 1990). Database information probably represents the most accurate information on drug utilization that can be obtained in elderly patients (Tamblyn et al. 1995). In drug dispensation databases, fairly accurate information can be obtained on drug intake that occurred a long time ago. Exposure information is available also for patients who are deceased or too ill to answer questions, without having to rely on proxy information. There is no potential for recall or interviewer bias which is always of concern with primary data collection. (3) Because the data are collected in an ongoing manner as a by-product of health-care delivery, epidemiological studies can be undertaken in reasonable time and at relatively low cost. The study variables are already available in computerized form and need not be obtained in time-consuming and expensive processes of data collection. (4) Some databases provide population-based data which cover the entire population of a geographical region and are thus fully representative of the population. Database studies do not require an informed patient consent and are therefore less prone to selection bias which may be a consequence of a low response rate in the study population.

Use of computerized databases for pharmacoepidemiology research is, however, not undisputed (Shapiro 1989), and there are a number of important limitations. A major concern is related to the validity of the diagnostic information contained in the database. In administrative health databases, diseases are primarily coded for billing and not for research purposes. There is no incentive for the health-care provider to use specific codes, e.g., "duodenal ulcer with bleeding" instead of "upper gastrointestinal bleeding otherwise not specified." Diseases are often coded according to the International Classification of Diseases (ICD-9 or ICD-10 coding schemes), and many different ICD-codes may be compatible with the same disease process. A combination of several diagnostic codes into a single "broader" code may therefore be necessary (Garbe et al. 1997, 1998a). Strom and Carson (1990) have described this problem by stating that researchers using diagnostic codes in a computerized database must be "lumpers" rather than "splitters."

The validity of diagnostic coding also depends on the ability of a diagnostic code to rather selectively represent the condition in question and therefore varies with the condition. Strom conducted a validation study of ICD-9 coding of Stevens–Johnson Syndrome in the COMPASS Medicaid database in the US (Strom et al. 1991). Records of 3.8 million patients in five US states were searched for ICD-9-CM code 695.1 which codes for Stevens–Johnson syndrome, but also for several other, less serious conditions. In an expert medical record review, only 14.8% of patients with ICD-9-CM code 695.1 whose medical records could be reviewed were judged to have Stevens–Johnson syndrome. Thus, studies of Stevens–Johnson syndrome in databases with ICD-9 coding cannot be conducted without additional validation of the diagnosis. Whenever possible, validity of disease coding should be quantified for each condition studied.

Validation studies usually use the paper medical or pharmacy record as the "gold standard." Access to patient charts for validation purposes is often obtained via the scrambled social security (patient identification) number. All personal identifying information is removed before copies of the charts are made available. The validation study by Strom also illustrates another problem: Only 51% of the medical records that were sought for in the study could actually be obtained (Strom et al. 1991). The authors state several reasons why they did not obtain access to the medical records: refusal of hospitals (30%), transcription errors (27%), translation of ID number not possible (17%), no location of medical record possible (22%), and other reasons (4%).

Administrative databases usually contain information on large numbers of patients; however, the amount of information per patient is limited: (1) Information about disease severity and many important risk factors that physicians use to make prescribing decisions is mostly lacking, and it may not be possible to exclude confounding by disease severity (Schneeweiss and Avorn 2005). In some instances, it is possible to construct an index of disease severity based on the patient's pharmacotherapy (Spitzer et al. 1992). (2) Relevant other confounder information for the association under study may not be contained in the database, creating a potential for bias. For example, most administrative databases do not contain information on smoking or alcohol use (Friedman et al. 2000) or age at menopause and reproductive history in women. (3) Administrative databases usually do not contain data on laboratory values or clinical measurements, although, in some databases, linkage with laboratory files has now become possible. (4) If relevant confounder information is missing and additional data collection required, it will make a study considerably more expensive and time-consuming, thereby diminishing some of the advantages connected with database research. It has to be decided on a case-by-case basis whether the information contained in a database is sufficient for the investigation of an association of interest and how much time and cost would be incurred if additional data have to be collected.

Although the medication information in databases is one of their major strengths, it also has some limitations: Information on drugs bought over the counter (OTC) and not prescribed by a physician is not available in the database; the patient's compliance with the prescription is unknown; in-hospital medication is usually not contained in the database; the prescribed daily dose is not documented in most databases, and the average daily dose has to be calculated instead based on the duration of drug use and the quantity of drug prescribed; the prescription file does not contain the indication for drug prescribing; however, in many instances, this information may be deduced from diagnostic coding in the ambulatory physician file; medication data will be truncated, if the database does not exist long enough or the database only includes elderly subjects. Truncation will limit the study of cumulative drug toxicity. Confounding by previous drug use may be avoided if a prior period of follow-up in the database is defined and the risk is only investigated in new users of the drug (Garbe et al. 1998b); many of the newest and/or most expensive drugs may not be available for study, if they are not included on the drug formulary.

Other important issues include whether the population contained in a database is representative of the source population and stable over time. For example, Saskatchewan Health Databases, which cover the population of the whole province of Saskatchewan, consist of a representative and fairly stable database population (Downey et al. 2003). In contrast, Medicaid databases in the USA are not representative of the US population, since they only include social welfare recipients and thereby over represent children, females, and non-whites in comparison with the total US population (Strom and Carson 1990). The skewness of Medicaid databases may not compromise the internal validity of a pharmacoepidemiology study conducted with these databases, but it may be a serious threat to its external validity, particularly when the data are being used for drug utilization research. In Medicaid Databases, turnover of the database population is high due to changing eligibility for Medicaid. Over a 5-year time period, only 35% of Michigan and 38% of Tennessee Medicaid enrollees were still in the system, with loss of eligibility being greatest in children and young adults (Ray and Griffin 1989). High patient turnover may also make it difficult to locate patient files for validation studies as has been reported in the study by Strom et al. (1991). Data from health maintenance databases are also not fully representative of the US population. Members of these organizations tend to be less frequently black or poor and have higher educational achievements. Turnover in membership at HMOs is usually less than in Medicaid databases (Friedman et al. 2000; Saunders et al. 2000).

48.5 Methodological Approaches for Pharmacoepidemiology Studies

The strategies employed to verify hypotheses on drug risks or benefits are similar to those used in other fields of epidemiology. The case-control design is the design of choice for the investigation of rare drug risks, particularly if multiple countries are necessary to attain sufficient power, while the cohort approach is preferably used to assess the risk of more frequent events or if more than one outcome has to be considered simultaneously. The special nature of drug exposure and the availability of existing databases in pharmacoepidemiology have given rise to specific challenges and preferred solutions to estimate risk and benefit.

48.5.1 Case-Control Studies

Field studies that employ a case-control design (cf. chapter ▸Case-Control Studies of this handbook) are not common in the evaluation of the risks and benefits of prescribed drugs. Drugs available over the counter without prescription are not recorded systematically in computerized databases and can therefore not be studied with linked databases. Thus studies of drugs such as analgesics, vitamin supplements, and anorexiants can only be studied by directly obtaining exposure information from subjects. With this design, cases with the outcome under study are identified from a given population, usually in hospitals or specialized clinics since they mostly involve serious outcomes. The approach to select the controls varies across studies. The International Agranulocytosis and Aplastic Anemia Study (IAAAS) that evaluated the effect of different analgesics on the risks of agranulocytosis and aplastic anemia used hospital-based controls (Anonymous 1986). The International Primary Pulmonary Hypertension Study (IPPHS) of the risk of primary pulmonary hypertension associated with anorexiant agents used patients treated by the same physician as the source of controls (Abenhaim et al. 1996). The Yale Hemorrhagic Stroke Study assessing the risk of diet and cough/cold remedies containing phenylpropanolamine used population-based controls identified by random-digit dialing (Viscoli et al. 2001) (see also chapter ▸Epidemiological Field Work in Population-Based Studies of this handbook). Finally, the transnational study on oral contraceptive risks used both hospital and population-based controls (Spitzer et al. 1996). Such field studies that collect information from patients, physicians, or medical charts are the exception because of the resources, expense, and time required to complete the study.

Case-control studies using existing health databases are much more common. Besides the usual concerns with case-control studies in general, some are specific to studies conducted from databases. For example, the General Practice Research Database (GPRD) was used to evaluate the impact of inhaled corticosteroids on the risk of hip fracture (Hubbard et al. 2002). The entire GPRD was used to identify 16,341 cases of hip fracture and a random sample of 29,889 subjects selected as controls. This design is attractive because of its efficiency in using only a sample of subjects to estimate an effect for an entire population. Such an approach, however, can be deceiving for specific diseases such as asthma because of the illusion of large sample sizes. Indeed, all cases of hip fracture selected within a population such as the GPRD suggest a very large sample size for the study, along with the very large number of controls. However, in assessing the effect of inhaled corticosteroids, a drug pertinent exclusively to the population of asthma or chronic obstructive pulmonary disease (COPD) patients, a large proportion of the cases and the controls are in fact irrelevant to the question at hand. Thus, for example, for the GPRD case-control study of hip fracture risk, only 878 of the 16,341 cases and 1,335 of the 29,889 controls were subjects with asthma or COPD (Hubbard et al. 2002). With its 16,341 cases and 29,889 controls, the study appears at first more powerful than the Quebec study, based on 3,326 cases of hip fracture and 66,237 controls selected from the asthma/COPD population (Suissa et al. 2004). In fact, the inference on the effect of inhaled corticosteroids is actually based on many fewer subjects than believed. This point is particularly relevant for studies that find no significantly increased risk since, despite appearances, conclusions are based on fewer numbers of cases and controls with respiratory disease and thus lower power than expected. Furthermore, the estimate of effect may be biased if the outcome of interest, in this instance fractures, is also associated with the disease itself (asthma or COPD) and not only with the medications used to treat these conditions. In this case, the bias due to the association with the disease can only be eliminated by restricting the analyses to the population of patients who have the disease (Suissa et al. 2004).

Another issue in such database case-control studies is the manner by which controls are selected and particularly the index date for controls. In the GPRD study of the impact of inhaled corticosteroids on the risk of hip fracture, controls were matched to cases on age, sex, general practice, and date of entry into the database (Hubbard et al. 2002). While the index date from which exposure was assessed was the fracture date for the cases, the index date for the controls was the same date as the matched case. To allocate such a date, one must be assured that the control is at risk on that date. Indeed, there could be control subjects with the same age, sex, general practice, and date of entry, but who are dead or not in the practice at the time of the matched case's fracture. These will be necessarily currently "unexposed," which could bias upward the rate ratio.

48.5.2 Cohort Studies

Observational database studies that use a cohort design (cf. chapter ▶Cohort Studies of this handbook) differ primarily with respect to their definition of cohort entry or time zero. The Saskatchewan asthma cohorts have defined asthma as well as its onset by the dispensing of medications used to treat the condition, without the use of diagnostic codes from physicians (Blais et al. 1998b; Suissa et al. 2002). Patients were considered to have asthma as of the first time they received three prescriptions for an asthma medication, including bronchodilators, inhaled steroids, and other asthma drugs, on at least two different dates within a 1-year period. The date of the third prescription defined the onset and diagnosis of asthma, and patients were then followed from that point on for the occurrence of asthma outcomes. Such a definition is not entirely accurate for two reasons: Subjects with asthma may be hospitalized at their initial presentation, and medications for asthma are used for other conditions such as COPD. In an attempt to exclude patients with COPD, age criteria were used, including only patients to the age of 44 and also excluding oral corticosteroids as one of the defining drugs for asthma.

Alternatively, cohort entry may be defined by calendar time. For example, a cohort formed from a health maintenance organization in eastern Massachusetts defined cohort entry as October 1, 1991 (Adams et al. 2002; Donahue et al. 1997). This cohort of 16,941 asthma patients was followed from this date or registration in the insurance plan to September 30, 1994. Such calendar time-based definitions of cohort entry will inherently define cohorts with patients who have varying durations of disease at time zero (cohort entry). Such a "prevalent" cohort, to be distinguished from an "incident" cohort defined by patients with new onset of asthma, can be subject to serious biases when evaluating the association between drug use and asthma outcomes. Indeed, if the risk of the asthma outcome and of being dispensed the drug under study are both associated with the duration of asthma, such prevalent cohorts will produce biased estimates of this association unless the duration of the condition can somehow be adjusted for. A source of selection bias for such prevalent cohorts is that the treatment itself may change because of prior events that are not included in the period of observation. For example, if a patient was hospitalized for asthma in the past and, as a result, was prescribed inhaled corticosteroids, such a patient may be at increased risk of a further hospitalization and of being dispensed inhaled corticosteroids subsequently. For such studies to be valid, information on the history of asthma prior to cohort entry, which includes the duration of the disease and prior outcomes such as asthma hospitalizations as well as prior drug exposures, are required for the purpose of adjustment or for testing for effect modification. A frequent problem with computerized database studies is that these historical data on the duration and history of the disease before cohort entry are rarely available.

The third type of cohort defines cohort entry by a specific clinical event, such as hospitalization, emergency room visit, or a physician visit. Here again, these cohorts can be incident or prevalent if these cohort-defining events are either the first one ever or rather the first to occur after a certain date. An example of this approach is a study from the Saskatchewan databases of asthma patients, with entry defined by the first time they received three prescriptions for an asthma medication within a 1-year period, after a 2-year span with no asthma medications. The study cohort consisted of all subjects hospitalized for asthma for the first time after cohort entry and followed until readmission. The use of inhaled corticosteroids subsequent to the first hospitalization was evaluated with respect to the rate of readmission (Blais et al. 1998a). A similar cohort definition was used with the Ontario database, although this cohort was based on elderly COPD patients and the COPD hospitalization defining cohort entry was not necessarily the first one to occur in their disease (Sin and Tu 2001a).

48.5.3 Nested Case-Control Studies

The complexity in data analysis is greater in the field of database studies because of the technical challenges presented by their large size. Indeed, the asthma cohort formed from the health maintenance organization in eastern Massachusetts included 742 asthma hospitalizations occurring during the 3 year follow-up period (Donahue et al. 1997). With over 16,000 patients in the cohort, an analysis based on the Cox proportional hazards model with time-dependent exposure (cf. chapter ▸Survival Analysis of this handbook) would require 742 risk sets (all patients in the cohort on the day of hospitalization), each containing approximately 16,000 observations with information on exposure and confounding factors measured at the point in time when the case occurred. Such an analysis would therefore require to generate close to 12 million observations (742 × 16,000), each with dozens of variables. Another example is the cohort study that included over 22,000 elderly patients hospitalized for COPD in Ontario, Canada, of whom around 8,000 either died or were re-admitted for COPD (Sin and Tu 2001a). A proper time-dependent analysis could include up to 140 million observations, creating a serious technical challenge in statistical computing. As a result of this complexity, the temptation to analyze these cohorts with exposures that are assumed not to change over time is attractive but, as described below, can cause severe immortal time bias (see Sect. 48.6.1).

Rather than analyzing such cohort studies with proper but complex time-dependent techniques, methods based on sampling can produce practically the same results at greater efficiency. The nested case-control design (cf. chapter ▸Modern Epidemiological Study Designs of this handbook), nested within the cohort, is precisely such an approach (Essebag et al. 2003; Suissa 2000). It is based on using data on all the cases with the study outcome that occurs during cohort follow-up. These represent the case series. A random sample of person-moments, namely, time points during the subjects' follow-up is then selected from all person-moments in the cohort to provide the control group for the nested case-control approach. For the cases, the index date, on which the timing of the exposure to the drug of interest is based, is simply the time at which the outcome occurred. For controls, the index date is the random person-moment(s) selected for that subject during follow-up, or the same point in time of the corresponding case (Suissa 2000). Because of the highly variable nature of drug exposure over time, it is important that person-moments are selected properly from all person-moments of follow-up for all members of the cohort. Thus, a subject may be selected more than once at different moments of their follow-up, and particularly person-moments preceding the index date of a case are valid control person-moments. For practical reasons and to conform to the Cox proportional hazards model, person-moments are usually selected from the risk set of each case. This approach involves identifying, for each case, all subjects who are at risk of the event at the time that the case occurred (the risk set) and controls are selected from this risk set (incidence density sampling, cf. chapters ▸Case-Control Studies and ▸Modern Epidemiological Study Designs of this handbook). Part of the simplicity of this approach is that all subjects in a risk set are allocated the same index date as the set-defining case.

The advantage of this approach is the direct relationship between the Cox proportional hazard model with time-dependent exposure and the conditional logistic regression analysis (cf. chapter ▸Regression Methods for Epidemiological Analysis of this handbook) that is used to analyze such nested case-control data. Thus, instead of using exposure data on all members of the risk set, as the Cox model would require, data on only a few subjects (usually 4 or 10 controls per case) are sufficient to provide a very efficient estimator of the rate ratio. Such ease of data analysis with the large size databases that are used in pharmacoepidemiology is crucial. As an example, with the asthma cohort study of Donahue, if 10 controls per case were used for each of the 742 cases, the analysis would be based on 7,420 observations instead of the 12 million observations necessary with the Cox model analysis.

In studying the effectiveness of drug treatment, one of the major problems is confounding by indication. The nested case-control approach becomes more useful, as it allows cases and controls to be matched on several measures of disease severity. Thus, the effect of a drug can be isolated, independently of the effects of the severity markers. Such a matched nested case-control study was used to evaluate the effectiveness of inhaled corticosteroids on asthma death (Suissa et al. 2000a). In that study, cases were identified from a cohort of over 30,000 asthma patients, from which 66 died of asthma. The Cox analysis would have required almost two million observations to be processed. For each case, however, the only members of the risk sets that were identified as controls were those with the same disease severity characteristics as the case, namely, prior hospitalization for asthma, oral corticosteroid use, number of canisters of beta-agonists, use of theophylline, and nebulized beta-agonists. Thus, cases and controls were similar on all these severity markers, except with respect to inhaled corticosteroids. As a result, the effect of inhaled corticosteroids could be assessed independently of these potential confounding factors.

48.5.4 Case-Crossover Design

Pharmacoepidemiology is frequently faced with the assessment of the risk of rare acute adverse events resulting from transient drug effects. Although the case-control approach can be used, the acuteness of the adverse event and the length of the drug's effect, as well as difficulties in determining the timing of drug exposure, induce uncertainty about the proper selection of controls. Moreover, confounding by indication may often be a problematic issue in such a design. In this situation, within-subject approaches have been proposed, including the case-crossover design and its extension the case-time-control design which was devised to counter time trend biases. The principle is that, when studying transient drug effects and acute outcome events, the best representatives of the source population that produced the cases are the cases themselves (cf. chapter ▸Modern Epidemiological Study Designs of this handbook).

To carry out a case-crossover study, three critical points must be considered. First, the study must necessarily be dealing with an acute adverse event which is alleged to be the result of a transient drug effect. Thus, drugs with regular patterns of use which vary only minimally between and within individuals are not easily amenable to this design. Nor are latent adverse events which only occur long after exposure. Second, since a transient effect is under study, the effect period (or time window of effect) must be precisely determined. An incorrect specification of this time window can have important repercussions on the risk estimate. Third, one must obtain reliable data on the usual pattern of drug exposure for each case, over a sufficiently long period of time.

The case-crossover study is simply a crossover study in the cases only. The subjects alternate at varying frequencies between exposure and non-exposure to the drug of interest, until the adverse event occurs, which happens for all subjects in the study, since all are cases by definition. With respect to the timing of the adverse event, each case is investigated to determine whether exposure occurred within the predetermined effect period. In the VACCIMUS study of Hepatitis B vaccination and the risk of a multiple sclerosis relapse, spontaneous reports indicated that such an effect could occur within 2 months of the vaccination (Confavreux et al. 2001). Thus, the case-crossover design used as the risk period the 2-month period prior to the onset of the relapse and any vaccination in this period to determine exposure status. To obtain control exposure, data on the average drug use pattern are necessary to determine the typical probability of exposure to the time window of effect. This is done by obtaining data for a sufficiently stable period of time prior to time of the event occurrence and its exposure period. For the VACCIMUS study, there were four control periods consisting of the four 2-month periods prior to the 2-month risk period. The estimation of the odds ratio is based on any appropriate technique for matched data (4 controls per case), such as conditional logistic regression.

This design has been used in pharmacoepidemiology (Barbone et al. 1998; Confavreux et al. 2001; Etienney et al. 2003; Fagot et al. 2001; Ki et al. 2003; Neutel et al. 2002; Sturkenboom et al. 1995).

48.5.5 Case-Time-Control Design

One of the limitations of the case-crossover design, particularly in the context of drug exposures, is that the exposure pattern may have changed over time and particularly between the control and risk periods. For example, a rapid increase in vaccination rates over time during the span of the case ascertainment for the VACCIMUS study, particularly if this span had been short, would have biased the estimate of the odds ratio. Indeed, this estimate would also include the effect of the natural time trend in exposure. If control subjects are available, the case-time-control design can be used to separate the time effect from the drug effect (Suissa 1995). In simple terms, the time effect is estimated from the case-crossover odds ratio of exposure among the control subjects. The net effect of exposure on event occurrence is then computed by dividing the combined time and drug effect estimated from the case-crossover odds ratio of exposure among the case subjects by the time effect (cf. chapter ▶Modern Epidemiological Study Designs of this handbook).

The approach is illustrated with data from the Saskatchewan Asthma Epidemiologic Project, a study conducted to investigate the risks associated with the use of inhaled ß-agonists in the treatment of asthma. Using databases from Saskatchewan, Canada, a cohort of 12,301 asthmatics was followed during 1980–1987. All 129 cases of fatal or near-fatal asthma and 655 controls were selected. The amount of ß-agonist used in the year prior to the index date, namely, high (more than 12 canisters per year) compared with low (12 or less canisters), was found to be associated to the adverse event. Of the 129 cases, 93 (72%) were high users of ß-agonists, compared with 241 (37%) of the 655 controls. The resulting crude odds ratio for high ß-agonist use is 4.4 (95% confidence interval (CI): 2.9–6.7). Adjustment for all available markers of severity, such as oral corticosteroids and prior asthma hospitalizations as confounding factors, lowers the odds ratio to 3.1 (95% CI: 1.8–5.4), the "best" estimate one can derive from these case-control data using conventional tools (Spitzer et al. 1992).

The use of inhaled ß-agonists, however, is known to increase with asthma severity which also increases the risk of fatal or near-fatal asthma. It is therefore not possible to separate the effects of the drug to the risk from that of disease severity, so that a within-subject design may be preferable. To apply the case-time-control design, exposure to ß-agonists was obtained for the 1-year current period and the 1-year reference period. Among the 129 cases, 29 were currently high users of ß-agonists and were low users in the reference period, while 9 cases were currently low users of ß-agonists and were high users previously. The case-crossover estimate of the odds ratio (*OR*) is thus 29/9 ($OR = 3.2$; 95% CI: 1.5–6.8). However, the high use of ß-agonists may have increased naturally over time, so that the control subjects were used to estimate this effect. Among the 655 controls, 65 were currently high users of ß-agonists and were low users in the reference period, while 25 were currently low users of ß-agonists and were high users previously, for an odds ratio of the time trend of 65/25 ($OR = 2.6$; 95% CI: 1.6–4.1). The case-time-control odds ratio, using these discordant pair frequencies for a paired-matched analysis, is given by $(29/9)/(65/25) = 1.2$ (95% CI: 0.5–3.0). This estimate, which excludes the effect of unmeasured confounding by disease severity, indicates a minimal risk for these drugs (Spitzer et al. 1992).

The case-time-control approach provides a useful complement to the case-crossover design when the probability of drug exposure is not stable over time, particularly between the control and risk periods (Donnan and Wang 2001; Hernandez-Diaz et al. 2003). However, its validity is subject to several assumptions, including the homogeneity of the odds ratio across subjects (Greenland 1996; Suissa 1998). This approach has been applied more frequently in pharmacoepidemiology recently (Corrao et al. 2005; Hernandez-Diaz et al. 2003; Kjaer et al. 2007; Nicholas et al. 2012; Park-Wyllie et al. 2009; Risselada et al. 2011; Schneider et al. 2005; Wang et al. 2012; Zambon et al. 2009).

48.5.6 Self-Controlled Case Series Method

The self-controlled case series (SCCS) method was first introduced in 1995 to study associations between vaccines and acute adverse events (Farrington 1995; Farrington et al. 1995). It gained broader attention when application of the method did not confirm the widely debated association between mumps, measles and rubella vaccines, and autism (Taylor et al. 1999). It has since been used more widely in vaccine, but also non-vaccine safety studies (Hubbard et al. 2003; Weldeselassie et al. 2011). The method was originally developed for the study of transient exposures with acute outcomes using only cases, i.e., subjects who had developed the outcome of interest (Whitaker et al. 2006), but the method was later extended also to non-acute outcomes, e.g., autism. The SCCS method estimates the relative incidence of the outcome of interest in a predefined postexposure risk period, compared to other times, which constitute the control period.

First, an overall study time window, usually defined by age or calendar time boundaries, is chosen in a way that the chance that individuals experience both risk and control periods is maximized. Then, individuals with the outcome of interest (cases) within this study time window are identified. For each case, the observation period – i.e., the time spent within the study time window – is determined. In a next step, the exposure histories of the cases are ascertained. The exposure dates of each case are used to define one or more risk periods, during which individuals are hypothesized to be at increased risk of the outcome of interest after the exposure. All other time within an individual's observation period that does not fall within a risk period is included in that individual's control period. To take account of only sampling cases, the likelihood is conditional on an event having occurred during the observation period. Having conditioned on the event in the observation period, and having fixed the risk and control intervals that each individual has progressed through, the only quantity that remains undetermined is the interval – i.e., the risk period or control period – in which that individual's event actually occurred. Since only this information contributes to the estimation of the exposure effect, an individual with zero events, i.e., with no information on when events occurred, does not contribute to the estimation and needs not be sampled. The relative incidence, i.e., the incidence in a risk period relative to the control period, is estimated by fitting a conditional Poisson regression model. Since inference is within individuals, the method implicitly adjusts for all confounders that remain fixed over the observation period, such as genetic or socioeconomic factors. Time-varying confounding factors, such as age, can be accounted for by subdividing each individual's observation period into age categories, which are modeled explicitly. The tutorial by Whitaker provides full practical details and worked examples (Whitaker et al. 2006). The SCCS method relies on three key assumptions: These are (1) events arise in a non-homogeneous Poisson process, (2) the occurrence of an event must not alter the probability of subsequent exposure, and (3) the occurrence of the event of interest must not censor or affect the observation period (Whitaker et al. 2009).

48.6 Some Methodological Challenges

48.6.1 Valid Drug Exposure Information and Relevant Exposure Risk Window

Accurate and complete exposure information is a prerequisite for both drug utilization and drug risk studies in pharmacoepidemiology. Data on drug exposures can be ascertained from different sources including exposures ascertained from self-report, medical records, electronic prescription data, or pharmacy dispensation data. Self-reported exposure data have in theory the advantage that drugs used many years previously can be ascertained, whereas this may not be the case in medical records or claims databases when a person's data do not go back long enough in the database or the database itself has not sufficient follow-up. Medications provided by self-report are more likely to reflect actual drug consumption than medications ascertained in databases, where information on patient compliance is lacking (West 1997). Collection of self-reported exposures in standardized interviews is, however, very expensive, time-consuming, and may be subject to interviewer bias. Moreover, recall inaccuracies will lead to exposure misclassification which is particularly of concern in studies of past drug exposures. Only few studies have investigated this issue and shown recall of past drug exposures to be influenced by the type of drug and the frequency of its administration (Behr et al. 2012; van den Brandt et al. 1991; West et al. 1995). A study which compared self-reported use of non-steroidal anti-inflammatory drugs (NSAIDs) and non-contraceptive estrogens against pharmacy dispensation data found that of those with only a single NSAID dispensation, 41% were able to recall *any* NSAID use compared with 85% for those with multiple NSAID dispensations, but only 30% recalled the NSAID name and only 15% the dose. For non-contraceptive estrogens, recall was higher with 78% of users recalling the name, but only 26% recalling the name and dose (West et al. 1995).

Drug exposure information ascertained from medical records has equally been shown to be less complete than that of pharmacy dispensation data (West et al. 1994). Even when a drug is suspected to cause an adverse drug reaction which led to hospitalization, it may not be entered in the inpatient medical record. Strom et al. found that for 128 hospitalized cases of Stevens–Johnson syndrome, only 50% of the 234 prescriptions for drugs suspected of causing the syndrome had been entered in the medical record when compared to the computerized Medicaid pharmacy claims files (Strom et al. 1991).

Since claims databases do not rely on physician's prescription documentation nor on patient recall, the drug exposure data from these databases are usually regarded as most complete. Completeness may, however, depend on the type of drug: If the medication of interest is rather inexpensive to purchase out-of-pocket compared to a person's drug insurance co-payment, this might decrease completeness and increase the potential for exposure misclassification. If OTC drug use or drugs administered during hospitalization are of relevance, medical record abstraction will often be necessary to obtain these types of drug exposures, since they are usually not contained in the databases. Medical records databases contain only the entry of the drug prescription, but it is unknown if the prescription was filled at a pharmacy. For this reason, they are more prone to misclassification of exposure than claims databases due to primary non-compliance, i.e., patients not filling the prescribed medication in a pharmacy. This issue has been investigated in several studies, with primary non-compliance ranging from 1.6% to 22% depending on the study and on the type of drug (Ekedahl and Mansson 2004; Esposito et al. 2008; Kennedy et al. 2008; Kirking et al. 2006).

Whereas information on the prescribed daily dose and the duration of each prescription is usually not contained in claims databases, this can be calculated in medical record databases based on the number of tablets to be taken per day

(Andersohn et al. 2009; Schade et al. 2007). This information is, however, entered in so many different ways in the medical record databases that it is tedious to extract and therefore underused. As a consequence, some medical record database studies did not calculate the duration of each prescription, but assigned fixed time windows for each prescription (Azoulay et al. 2010; Opatrny et al. 2008) or estimated the duration of exposure by the number of prescriptions written (Bodmer et al. 2012; Brauchli et al. 2011; Meier et al. 2010).

The correct definition of the relevant exposure to a medical drug for the outcome of interest (referred to as the exposure risk window) is a great challenge in all pharmacoepidemiology studies. For example, if an adverse outcome is known to occur only immediately after initial use of a drug and the exposure definition includes all of the patient's time on a drug in the analysis, a significant amount of non-relevant exposure time could be included and could produce biased risk estimates (van Staa et al. 1994). Information from other sources, e.g., data from spontaneous reporting systems or published case reports, can increase the likelihood of focusing on relevant periods of exposure. Sensitivity analyses might provide further insight into the relevant exposure risk window.

For drugs presenting an immediate hazard after initial use, an incident or new user cohort design has been recommended (Ray 2003). When defining new users, investigators have to define look-back periods to ensure that prevalent users are not incorrectly classified as new users, e.g., patients who were already using the drug of interest when they entered the database.

48.6.2 Immortal Time Bias in Cohort Studies

A challenge of cohort studies is in their data analysis. Since drug therapy, the exposure of interest, often changes over time, data analysis must take this variability into account. However, such variability in exposure over time is not simple to incorporate in the analysis. Due to the complexity of such analyses, several of the studies mentioned above employed a time-fixed definition of exposure, by invoking the principle of intention-to-treat analysis. This principle, borrowed from randomized controlled trials, is based on the premise that subjects are exposed to the drug under study immediately at the start of follow-up. This information is unknown in database studies.

To emulate randomized controlled studies in the context of cohort studies, some authors have looked forward after cohort entry for the first prescription of the drug under study. In this way, a subject who was dispensed a prescription for such drug was considered exposed and a subject who did not was considered unexposed. Different time periods of exposure assessment were used. For instance, in the context of COPD, a prescription for inhaled corticosteroids during the period of 90 days after cohort entry was used to define exposure (Sin and Tu 2001b). In other studies, periods of 1 and 3 years were used to consider subjects exposed to inhaled corticosteroids in assessing their impact on mortality (Sin and Man 2002;

Sin and Tu 2001a). This approach, however, leads to immortal time bias, a major source of distortion in the rate ratio estimate (Suissa 2003).

Immortal time bias arises from the introduction of immortal time in defining exposure by looking forward after cohort entry. Indeed, if exposed subjects were classified as such because they were observed to have been dispensed their first prescription for an inhaled corticosteroid 80 days after cohort entry, they necessarily had to be alive on day 80. Therefore, this 80-day period is immortal. While some exposed subjects will have very short immortal time periods (a day or two), others can have very long immortal periods. On the other hand, unexposed subjects do not have any immortal time, and in particular the subjects who die soon after cohort entry, with too little time to receive the drug under study. Therefore, the exposed subjects will have a major survival advantage over their unexposed counterparts because they are guaranteed to survive at least until their drug was dispensed.

This generation of immortal time in exposed subjects, but not in the unexposed subjects, causes an underestimation of the rate of the outcome among the exposed subjects. This underestimation results from the fact that the outcome rate in the exposed is actually composed of two rates. The first is the true rate, based on the person-time cumulated after the date of drug dispensing that defines exposure (post-*Rx*), while the second is that based on the person-time cumulated from cohort entry until the date of drug dispensing that defines exposure (pre-*Rx*). The first rate will therefore be computed by dividing all outcome events in that group by the first rate person-time, while the second rate will by definition divide zero events by the second rate person-time. For example, the rate in the exposed

$$\text{rate} = \text{deaths}/\text{total person-years}$$

consists in fact of two rates:

$$\text{rate pre-}Rx = 0/\text{person-years pre-}Rx$$

and

$$\text{rate post-}Rx = \text{deaths}/\text{person-years post-}Rx.$$

The zero component of the rate will necessarily bring down the exposed rate. Since there is no such phenomenon in the unexposed group, the computation of the rate ratio will systematically produce a value lower than the true value because of the underestimation of the exposed rate. In particular, if the drug under study is altogether unrelated to the outcome, so that the true rate ratio is 1, this approach will produce rate ratios lower than 1, thus creating an appearance of effectiveness for the drug.

The immortal time in exposed subjects also causes an overestimation of the rate of the outcome among the unexposed subjects. This is because the zero component of the rate in the exposed group should in fact be classified in the unexposed group. Indeed, subjects are in fact unexposed to the drug under study between cohort entries until the date of drug dispensing that defines exposure. They only start to be exposed after the drug is dispensed. Thus, the zero rate should in fact be combined with the unexposed rate.

Immortal time bias is thus the result of simplistic yet improper exposure definitions and analyses that cause serious misclassification of exposure and outcome events. This situation is created by using an emulation of the randomized controlled trial to simplify the analysis of complex time-varying drug exposure data. However, such studies do not lend themselves to such simple paradigms. Instead, time-dependent methods for analyzing risks, such as the Cox proportional hazard models with time-dependent exposures or nested case-control designs, must be used to account for complex changes in drug exposure and confounders over time (Levesque et al. 2010; Samet 2003; Suissa 2003, 2004, 2007, 2008).

48.6.3 Confounding by Indication and Unmeasured Confounding

The indication for which a medication is given may act as a confounder in observational studies, particularly when assessing the effectiveness of a drug (Horwitz and Feinstein 1981; Slone et al. 1979; Strom et al. 1983). Such confounding by indication will be present if the indication for the prescription of the medication under study is also a determinant of the outcome of interest. Generally, a drug is more likely to be prescribed to a patient with more severe disease who, in turn, is more likely to incur an adverse outcome of the disease. Thus, patients prescribed the drug under study will have higher rates of outcome than the subjects not prescribed the drug. Such an appearance of lack of effectiveness could simply be a reflection of the effect of indication, in this case disease severity.

Confounding by indication is often difficult to control, primarily because the precise reason for prescribing is rarely measured. This may preclude the study of drug effectiveness with observational designs (Miettinen 1983). Yet, a clinical trial to answer this question would require the follow-up of thousands of patients over a long time, which may simply be unfeasible. Observational studies become the tool of choice as long as validity of the study is not compromised by intractable confounding by indication (Miettinen 1983). If such an observational study produces lower rates of outcome for the drug under study, one may conclude on the one hand that these medications are effective. On the other hand, if users of the drug are found to be at equal or increased risk of the outcome relative to non-users, it would not be possible to conclude on the absence of a protective effect of these medications.

This problem of confounding by indication is compounded with the use of computerized databases, because of their lack of information on important confounders (Shapiro 1989). The absence of information on drug indication precludes the control of confounding by adjustment in the analysis. Thus, control for confounding by indication must be tackled at the design level. One approach is to restrict the study to a group of patients homogeneous with respect to disease severity. For example, in a study of the effectiveness of inhaled corticosteroids in asthma, Blais et al. (1998a) identified a point in time at which users and non-users of inhaled corticosteroids would have a similar level of asthma severity. The study was thus restricted to patients who had just been hospitalized for asthma, with the discharge date taken as time zero, which would greatly reduce heterogeneity in disease severity. The rate of a readmission for asthma was then assessed according to the use of inhaled corticosteroids after this initial hospitalization.

Another approach is to compare two medications prescribed for the same indication (Strom et al. 1983, 1984). In this case, relative effectiveness as opposed to absolute effectiveness will be evaluated. An example of this approach was also used in a study of the effectiveness of early use of inhaled corticosteroids in asthma. Blais et al. (1998b) identified a cohort of newly treated asthma patients and compared regular users of inhaled corticosteroids with regular users of either antiallergic agents or theophylline and matched for the duration of asthma at the initiation of therapy.

In addition to disease severity, genetic factors can determine individual susceptibility to both dose-dependent and dose-independent ADRs.

More recently, instrumental variable (IV) methods have been proposed as a possible solution to confounding by indication or otherwise unmeasured confounding in comparative effectiveness and drug safety studies, provided a suitable instrument can be identified (Brookhart et al. 2006a, 2010; Rassen et al. 2009a, b). The idea is that the causal effect of exposure on outcome can be captured by using the relationship between the exposure and another variable, the IV (Martens et al. 2006).

For an IV to provide unbiased treatment effects even in the presence of unmeasured confounders, three important assumptions have to be fulfilled: (i) The IV should be correlated to the exposure under study, (ii) the relationship between the IV and the exposure should not be confounded by other variables, and (iii) the IV should influence the outcome neither directly nor indirectly by its relationship with other variables and should be associated with the outcome only by its association with the exposure. As such, an IV can be thought of as an observed variable that is associated with variation in the exposure similar to randomized assignment. If these assumptions are met, IV analyses can provide unbiased treatment estimates even in the presence of unmeasured confounding. If both the instrument and the exposure are dichotomous, the classic IV estimator, also called the Wald estimator, is given by

$$\beta_{\mathrm{IV}} = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[X|Z=1] - E[X|Z=0]}$$

where X is the treatment, Y is the outcome, Z is the instrument, and β is a measure of the effect of X on Y. The numerator of this estimator is comparable to the intention-to-treat (ITT) estimate of an RCT measured as a risk difference with the randomization providing Z. The denominator presents the difference in treatment rates between levels of Z. In an RCT, this would refer to the compliance in the different treatment arms. In the case of perfect compliance, i.e., when Z perfectly predicts X, then $E[X|Z=1] - E[X|Z=0] = 1$, the IV estimator will be identical to the ITT estimator (Brookhart et al. 2010).

It is, however, not possible to empirically verify assumptions (ii) and (iii) regarding the instrument making it difficult in practice to find a suitable one (Hernan and Robins 2006). In pharmacoepidemiology studies, where the instrument needs to be unrelated to patient characteristics, physician prescribing preference has been used in several studies (Brookhart et al. 2006b; Schneeweiss et al. 2007, 2008; Wang et al. 2005, 2007), but this instrument may actually have violated assumptions (ii) and (iii) leading to bias (Hernan and Robins 2006). In the case of a weak instrument, i.e., when the instrument is only weakly related to the exposure under study, this would actually increase the bias (Hernan and Robins 2006; Martens et al. 2006). It has also been pointed out that the standard IV methodology deals only poorly with time-varying exposures which are most frequently the case in pharmacoepidemiology studies (Hernan and Robins 2006).

If a valid instrument cannot be identified, sensitivity analyses may provide further insight into the impact of residual or unmeasured confounding. Basic sensitivity analyses are conducted to assess how strong and how imbalanced a confounder would have to be between the exposures of interest to explain the observed effect. A review of sensitivity analysis and adjustment for unmeasured confounders with external data in pharmacoepidemiology studies is provided by Schneeweiss (2006).

48.6.4 Depletion of Susceptibles

In general, the risk of an ADR associated with drug use does not remain constant over time, and can change in different ways from the start of its use. The risk may increase with cumulative drug exposure (e.g., the risk of cardiomyopathy associated with cumulative anthracycline exposure or the risk of cataract associated with continued glucocorticoid use), but it may also decrease after an initial period of sharp increased risk. Therefore, in using a case-control study that evaluates the effect of current use of a drug, past history of use of a drug or a class of drugs must be accounted for as it may modify the risk of an ADR associated with current use of the drug.

A decreasing risk after an initial period of increased risk is probably more important than cumulative drug toxicity and may, at the population level, lead to a phenomenon which has been described as “depletion of susceptibles”: Patients who remain on the drug are those who can tolerate it, while those who are susceptible to adverse drug reactions will stop the drug and thereby select themselves out of the exposed cohort (Moride and Abenhaim 1994). Such a pattern has been demonstrated for the gastrointestinal toxicity of non-steroidal anti-inflammatory drugs (NSAIDS). It has been shown that the risk of upper gastrointestinal bleeding (UGIB) was highest after the third NSAID prescription and thereafter decreasing (Carson et al. 1987). Moride and Abenhaim (1994) empirically showed a depletion of susceptibles effect in a hospital-based case-control study of NSAIDS and the risk of UGIB. They investigated the risk of UGIB associated with recent NSAID use stratified by past or no past NSAID use. The risk of UGIB was significantly greater for those patients who used NSAIDS for the first time in 3 years ($OR = 22.7$) than for those who had used these drugs before ($OR = 3.0$) (Moride and Abenhaim 1994).

The enormous importance of accounting for changes in drug risk over time in the design and/or analysis of pharmacoepidemiology studies was highlighted in the debate about the risk of venous thromboembolism (VTE) associated with second-or third-generation oral contraceptives (OC). Several pharmacoepidemiology studies published in 1995/1996 reported an increased risk of VTE among users of newer OC preparations compared with those of older OC preparations (Anonymous 1995b; Bloemenkamp et al. 1995; Jick et al. 1995; Spitzer et al. 1996). Additional analyses suggested that the magnitude of the risk estimates for individual OC was closely linked with the time of market introduction of the respective OC, with increasing risk for the newer preparations (Lewis et al. 1996). Since a larger proportion of users of older OC preparations were long-term users compared with those using newer OC, a depletion of susceptibles effect was postulated to be active within these studies. It was hypothesized that individuals with good tolerance were preferentially long term users of older OC preparations, whereas groups with shorter duration of use might be more frequently using the newer OC preparations and thereby constitute a different subpopulation.

The phenomenon of depletion of susceptibles can lead, if not accounted for properly, to a comparison of OC medications with different years of entry into the market and result in an overestimation of the risk associated with the most recently introduced medications. To properly account for depletion of susceptibles, the approach to statistical analysis must take account of the duration and patterns of OC use and of course have the available data to do so. In this example, the pattern and duration of OC use are not confounders, but effect modifiers of the risk of VTE associated with recent OC use. OC pattern and duration can therefore not be simply "adjusted for" in the statistical analysis, but a stratified analysis has to be conducted which compares the risk of VTE for the different OC preparations for the different durations and patterns of use. When the analysis was restricted to the same pattern of OC use, distinguishing between first time users, repeaters, and switchers, the risk of VTE as a function of the duration of oral contraceptive use was essentially the same for second-and third-generation pills relative to never users (Suissa et al. 1997, 2000b).

48.6.5 Use of Propensity Scores in Pharmacoepidemiology

In a randomized trial, randomization of study subjects to different treatment regimens aims to assure the absence of systematic differences between the patients in terms of measured and unmeasured confounders. In observational studies, direct comparisons of the outcomes of treated and untreated patients may be misleading because of systematic differences between those patients who have and have not received treatment. The propensity score has been proposed as a method of adjusting for covariate imbalances in an observational study and has recently been proposed also for pharmacoepidemiology research (Perkins et al. 2000;

Wang and Donnan 2001; Wang et al. 2001). The propensity score $\pi(X)\!:= Pr(\text{exposed}|\ X)$ is defined as the conditional probability of receiving a particular treatment (i.e., being exposed) given the set of observed covariates X. The propensity score thus represents a summary of the covariates X that are associated with treatment allocation in the form of a single variable.

The propensity score approach is a two-stage approach. At the first stage, the propensity score is estimated for each study subject based on the values of the observed covariates. The most commonly used approach is to obtain the propensity score estimates in a logistic regression model, where treatment allocation is used as the dependent (response) variable and observed potential confounders X are used as explanatory variables: $\log(\frac{\pi(X)}{1-\pi(X)}) = X \cdot \beta$, where the regression coefficients β are estimated by maximum likelihood.

Having obtained the estimated propensity score, it can be used by a number of approaches at the second stage: Study subjects may be matched or stratified based on their propensity scores or the propensity scores may be adjusted for in a regression model (Wang and Donnan 2001).

Many published applications use stratification by the propensity score. A common approach is to stratify by quintiles of the distribution of the estimated propensity scores and to test the balance of each confounder between the treatment groups in each stratum (Wang and Donnan 2001). Having patients with similar propensity scores in each stratum, it may be assumed that the covariate distributions in the two treatment groups are equally similar within each stratum, so that the treatment assignment within the strata can be functionally regarded as random. If unbalanced confounders are still found, the propensity score model may be re-estimated with modifications until balance is achieved.

Stratification by the propensity score cannot control confounder effects within a single stratum. The somewhat arbitrary choice of five strata can be viewed as a compromise between reduction of bias and robustness of the results: An increasing number of strata will reduce the bias in the stratification estimate, but it will at the same time decrease the robustness of the results when the sample size of the smaller arm in a stratum becomes too small. Control of bias through use of propensity scores is based on the following assumptions:

- All subjects must have some non-zero probability of receiving each treatment (referred to as the "strongly ignorable assumption"). This ensures independence of treatment assignment and response variable within propensity score strata.
- Treatment assignment depends solely on the observed covariates, i.e., *all* confounders are included in the propensity score model.

If *all* confounders were not ascertained or confounder measurement was associated with bias, the use of propensity scores will not eliminate bias. In fact, the study will be subject to the same bias as an observational study that did not measure all confounders and could only incompletely adjust for known confounders. The use of propensity scores may, however, help to detect incomparability between treatment groups (i.e., lack of overlap in covariate values) that may remain undetected in a standard regression model. It also provides an additional tool to assess the performance of the traditional regression model (Wang and Donnan 2001).

Propensity scores have more often been used in cohort studies of drug effectiveness (Mojtabai and Zivin 2003; Schroder et al. 2003; Seeger et al. 2003; Young-Xu 2003), but they have also been applied to studies of drug safety (MacDonald et al. 2003). Nevertheless, the use of propensity scores may be a particular challenge in the common situation in pharmacoepidemiology of time-dependent exposures and covariates. Moreover, with the very large sizes of databases used in pharmacoepidemiology, the need to reduce the number of covariates to a single score is not crucial, and thus, the advantage of propensity scores compared to including the confounders directly in the model of data analysis becomes less evident. Recently, the use of propensity scores has been extended to semiautomated covariate selection called high-dimensional (HD) propensity score to identify surrogates for unobserved confounder information (Rassen et al. 2011; Schneeweiss et al. 2009; Schneeweiss and Rassen 2011).

48.7 Drug Utilization Studies

Drug utilization studies are an important tool in improving rational drug use and providing data for cost/benefit considerations. Drug utilization has been defined as the “prescribing, dispensing, administering, and ingesting of drugs” (Serradell et al. 1991). This definition implies that several steps are involved in drug utilization and that, consequently, in each of these steps, problems in drug use can arise. The World Health Organization defines drug utilization in a broader sense as the “marketing, distribution, prescription and use of drugs in a society, with special emphasis on the resulting medical, social and economic consequences” (World Health Organization 1977), thereby including also the effects of drug use on the population. Apart from examining drug use, goals of drug utilization studies include the identification of problems of drug utilization with respect to their importance, causes, and consequences; the establishment of a scientific basis for decisions on problem solving; and the assessment of the effects of actions taken. Some examples for studies that illustrate these goals are the following: What are the prevalence, pattern, and risk factors of use for benzodiazepines in Italy? What is the quality of NSAID prescribing in Croatia and Sweden (Vlahovic-Palcevski et al. 2002)? Are labeled contraindications to the use of cisapride adhered to (Weatherby et al. 2001)? What is the impact of safety alerts on the prescribing of a drug (de la Porte et al. 2002; Weatherby et al. 2001)? After a drug has been withdrawn from the market, in which way does drug utilization of related drugs change (Glessner and Heller 2002)? What are characteristics of physicians and practices that make early use of new prescription drugs (Tamblyn et al. 2003)?

Drug utilization studies can be qualitative and quantitative. *Quantitative* drug utilization studies are conducted for a number of purposes: to ascertain the quantities of drugs consumed in a specific period and in a specific geographical area, to investigate the development of drug utilization over time, to compare and contrast the use of a drug between different geographical areas, to identify possible over- or underutilization of drugs, to determine trends in drug use according to population demographics, to estimate the prevalence of illness based on the consumption of drugs utilized in its treatment, and to compare the prevalence of an illness in different areas.

The main aim of *qualitative* drug utilization studies is to determine the appropriateness of drug prescribing. They require the a priori establishment of quality indicators against which drug utilization is compared. National or international expert panels are sometimes used to help defining quality indicators in a consensus process (McLeod et al. 1997). Quality indicators may be based on the following parameters: the medical necessity for drug treatment; adherence to labeling with respect to labeled indications, contraindications, or interactions; duration and dose of treatment; use of fixed drug combinations when only one of its components would be justified; availability of treatment alternatives which are more effective or less hazardous; availability of an equivalent less costly drug on the market; etc. In North America, these studies are known as drug utilization review (DUR) studies. DUR studies are aimed at detecting and quantifying problems of drug prescribing. They should be distinguished from DUR programs which are interventions in the form of an authorized, structured, and ongoing system to improve the quality of drug prescribing (Lee and Bergman 2000). In contrast to DUR studies which provide only minimal feedback to the involved prescribers and are not interventional by their design, DUR programs include efforts to correct inappropriate patterns of drug use and include a mechanism for measuring the effectiveness of corrective actions taken to normalize undesirable patterns of drug use (Hennessy and Strom 2000).

For quantitative and qualitative studies, it would be ideal to have a count of the number of patients who either ingest a drug of interest during a certain time frame or who use a drug inappropriately in relation to all patients who received the drug during a given time frame. The available data are often only approximations of the number of patients and may be based on cost or unit cost, weight, number of prescriptions written or dispensed, and number of tablets, capsules, doses, etc., sold (Lee and Bergman 2000). Drug cost data have a number of limitations, since the price of a drug is not the same within and across countries. Drug pricing may be affected by different drug distribution channels, the quantities of drugs purchased, exchange rate fluctuations, different import duties, and regulatory policies that affect pricing (Serradell et al. 1991). Studies based on the overall weight of a drug sold are similarly limited, since tablet sizes vary which makes it difficult to translate weight even into the number of tablets sold (Lee and Bergman 2000). The number of prescriptions written or dispensed for a particular drug product is a measure that is frequently used in drug utilization studies. However, the number of prescriptions for different patients in a given time interval varies and also the supply of drugs prescribed. To estimate the number of patients, one must divide by the average number of prescriptions per patient. The number of tablets, capsules, etc., sold is often used in conjunction with the defined daily dose (DDD) measurement unit for drug use.

The DDD is the assumed average maintenance dose per day for a drug used for its main indication in adults. Use of the DDD underlies two basic assumptions: that patients are compliant and that the doses used for the major indication are

Table 48.6 The structure of the ATC coding system for metformin

A	Alimentary tract and metabolism (first level, anatomical main group)
A10	Drugs used in diabetes (second level, therapeutic subgroup)
A10B	Oral blood glucose-lowering drugs (third level, pharmacological subgroup)
A10B A	Biguanides (fourth level, chemical subgroup)
A10B A02	Metformin (fifth level, chemical substance)

the average maintenance doses (Serradell et al. 1991). The DDD dosing levels are assigned per ATC fifth level by the WHO Collaborating Centre for Drug Statistics Methodology in Norway based on recommendations in the medical literature. The DDD provides a fixed unit of measurement independent of the price and formulation of a drug. It can be used to examine changes in drug consumption over time and permits international comparisons. The DDD is a technical unit of measurement and does not necessarily reflect the recommended or prescribed daily dose (PDD). Doses for individual patients and patient groups will often differ from the DDD based on individual patient characteristics such as age, weight, and pharmacokinetic considerations. DDDs may be used to obtain crude estimates of the number of persons exposed to a particular drug or class of drugs and are sometimes used as denominator data for crude estimation of ADR rates (Kromann-Andersen and Pedersen 1988; Leone et al. 2003). This use of the DDD methodology is rather limited in drugs with more than one indication, particularly when the drug dose differs for each indication. It is also limited when the duration of drug treatment varies greatly between patients. The DDD does not take into account pediatric use of a drug. DDDs are not established for topical preparations, sera, vaccines, antineoplastic agents, allergen extracts, general and local anesthetics, and contrast media.

The defined daily dose (DDD) is usually used in conjunction with the Anatomical Therapeutic Chemical (ATC) coding system. Coding for the ATC system at the WHO Collaborating Centre for Drug Statistics Methodology in Norway is based on requests from users including manufacturers, regulatory agencies, and researchers. In the ATC system, drugs are classified in groups at five hierarchical levels. The drugs are divided into 14 main groups (first level), with one pharmacological/therapeutic subgroup (second level). The third and fourth levels are pharmacological/therapeutic and chemical subgroups, and the fifth level is the chemical substance. Table 48.6 illustrates the structure of the ATC coding system using metformin as an example. In the ATC system, all plain metformin preparations are thus given the code A10B A02.

Coverage of the ATC system is not comprehensive: Complementary and traditional medicinal products are generally not included in the ATC system; ATC codes for fixed combination drugs are assigned only to a limited extent; some drugs may not be included in the system since no request for coding has been received by the WHO Collaborating Centre; a medicinal product that is used for two or more equally important indications will usually be given only one code based on its main indication which is decided from the available literature. On the other hand, a medicinal product can have more than one ATC code if it is available in two or more strengths or formulations with clearly different therapeutic uses.

Use of computerized databases has greatly facilitated drug utilization research. Databases can be distinguished into those which include both drug and diagnostic data (examples have been given in Sect. 48.4) and into those which include only drug data (e.g., Denmark's Odense Pharmacoepidemiologic Database, Denmark's Pharmacoepidemiologic Prescription Database of the County of North Jutland, Spain's Drug Data Bank, Sweden's County of Jämtland Project). Some more databases used in drug utilization research and a description of these databases can be found in the references (Lee and Bergman 2000; Serradell et al. 1991).

Interpretation of drug utilization data needs appropriate care. Observed geographical or time differences in drug utilization may be caused by many factors different from prescribing behavior, e.g., differences in the age and sex distribution, different patterns of morbidity, change in diagnostic criteria, and differences in the access to health care. If used appropriately, drug utilization research provides a powerful scientific tool to identify factors that influence drug prescribing and to develop strategies to modify prescribing behavior. Further research is needed to determine which characteristics of inappropriate prescribing are susceptible to modification and what are the most efficient intervention strategies.

48.8 Conclusions and Prospects

Pharmacoepidemiology is still a relatively young scientific discipline. Over the last 20–30 years, there has been enormous progress in the improvement of its methods and development of new approaches to studies of drug safety and effectiveness. Pharmacoepidemiology has taken advantage of the rapidly expanding methods in epidemiology and has developed sophisticated methods to cope with problems that are specific to the field. New statistical approaches have been developed for signal generation based on data from the spontaneous reporting systems. Large computerized health-care utilization databases are now widely used for research into beneficial and harmful drug effects, their use being facilitated by the development of more and more powerful computer technologies. With the experience gained through the use of these data and a careful understanding of the underlying health-care systems in which the data were generated, computerized databases provide a highly useful data source for pharmacoepidemiology studies. Despite the often enormous size of these databases, statistical power may be lacking for rare exposures or rare outcomes within a single database. In Europe and North America, there are several initiatives to combine data from multiple databases for pharmacoepidemiology studies. The European Medicines Agency has coordinated the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCEPP) since 2006 to strengthen post-authorization safety surveillance of medicinal products in Europe by facilitating the conduct of multicenter post-authorization safety studies across different European countries. In the USA, the Sentinel Initiative has been launched by the Food and Drug Administration in May 2008 to proactively monitor the safety of drugs, biologics, and medical devices and to provide a database for the rapid investigation of medicinal product safety issues. The Sentinel Initiative was set up with the goal to access data from 100 million patients from several data sources by July 2012. In Canada, the Canadian Network for Observational Drug Effect Studies (cNODES) was launched in October 2011 consisting of 60 researchers at universities across Canada who will draw data from existing databases representing about 27 million people. Combining data from multiple sources brings exciting new challenges for pharmacoepidemiologists with respect to issues of data privacy, coordination of research projects, governance and organizational issues, as well as methodological issues such as the combination and analysis of data with different data structure and data availability, the harmonization of terminologies when the data are coded according to different coding systems, and different approaches to finally synthesize the data. In Europe, experience on these issues is already gained in some multi-country database studies such as SOS (Safety of Non-steroidal Anti-Inflammatory Drugs), ARITMO (Arrhythmogenic Potential of Drugs), SAFEGUARD (Safety Evaluation of Adverse Reactions in Diabetes), and some other projects, some of which are part of the Innovative Medicines Initiative. In the USA, experience comes from the Mini-Sentinel pilot program which includes 25 participating institutions. In Canada, first experiences with the network are being obtained within a study on the renal safety of statins.

Regarding methods, there has been a progressive refinement of case-control and cohort studies, and efficient sampling strategies within a cohort are now often employed. The case-crossover and case-time-control designs are being used for the study of acute transient drug effects to eliminate control selection bias and confounding by indication or other factors. Propensity scores are increasingly used as a method to minimize confounding in the study of intended drug effects. High-dimensional propensity score methods have been proposed to improve confounder control by adjusting for surrogates of unmeasured confounders. Instrumental variable analyses provide another approach to unmeasured confounding, if a valid instrument can be found. Some databases, e.g., the UK CPRD, meanwhile offer initial randomization of patients with electronic follow-up of randomized patients in the database (“randomize into database epidemiology”). Drug utilization review programs are now required in all US hospitals and have been implemented voluntarily in many other health-care programs which will lead to further refinement of drug utilization research. A great challenge ahead is linkage of pharmacoepidemiology studies with the latest techniques of genetics, biochemistry, immunology, and molecular biology. The CPRD has over 300 practices with access to over 2 million patients that are willing to collect blood or other samples suitable for genetic research. It is of particular interest to understand why individuals respond differently to drug therapy, both in terms of beneficial and adverse effects. Investigation of the genetic makeup of study patients on a population level will be greatly facilitated through the enormous progress in pharmacogenomics and molecular biology. New study designs may emerge as a consequence of these developments. It remains to be explored to which extent database studies may be used to include molecular genetic or immunologic investigations.

References

Abascal VM, Larson MG, Evans JC, Blohm AT, Poli K, Levy D (1998) Calcium antagonists and mortality risk in men and women with hypertension in the Framingham Heart Study. Arch Intern Med 158:1882–1886

Abenhaim L, Moride Y, Brenot F, Rich S, Benichou J, Kurz X, Higenbottam T, Oakley C, Wouters E, Aubier M, Simonneau G, Begaud B (1996) Appetite-suppressant drugs and the risk of primary pulmonary hypertension. International Primary Pulmonary Hypertension Study Group. N Engl J Med 335:609–616

Adams RJ, Fuhlbrigge AL, Finkelstein JA, Weiss ST (2002) Intranasal steroids and the risk of emergency department visits for asthma. J Allergy Clin Immunol 109:636–642

Almenoff JS, Pattishall EN, Gibbs TG, DuMouchel W, Evans SJ, Yuen N (2007) Novel statistical tools for monitoring the safety of marketed drugs. Clin Pharmacol Ther 82:157–166

Amann U, Schmedt N, Garbe E (2012) Prescribing of potentially inappropriate medications for the elderly: an analysis based on the PRISCUS list. Dtsch Arztebl Int 109:69–75

Andersohn F, Schade R, Suissa S, Garbe E (2009) Long-term use of antidepressants for depressive disorders and the risk of diabetes mellitus. Am J Psychiatry 166:591–598

Anonymous (1986) Risks of agranulocytosis and aplastic anemia. A first report of their relation to drug use with special reference to analgesics. The International Agranulocytosis and Aplastic Anemia Study. JAMA 256:1749–1757

Anonymous (1995a) Venous thromboembolic disease and combined oral contraceptives: results of international multicentre case-control study. World Health Organization Collaborative Study of Cardiovascular Disease and Steroid Hormone Contraception. Lancet 346:1575–1582

Anonymous (1995b) Effect of different progestagens in low oestrogen oral contraceptives on venous thromboembolic disease. World Health Organization Collaborative Study of Cardiovascular Disease and Steroid Hormone Contraception. Lancet 346:1582–1588

Azoulay L, Schneider-Lindner V, Dell'Aniello S, Filion KB, Suissa S (2010) Thiazolidinediones and the risk of incident strokes in patients with type 2 diabetes: a nested case-control study. Pharmacoepidemiol Drug Saf 19:343–350

Barbone F, McMahon AD, Davey PG, Morris AD, Reid IC, McDevitt DG, MacDonald TM (1998) Association of road-traffic accidents with benzodiazepine use. Lancet 352:1331–1336

Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, De Freitas RM (1998) A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol 54:315–321

Bate A, Lindquist M, Edwards IR (2008) The application of knowledge discovery in databases to post-marketing drug safety: example of the WHO database. Fundam Clin Pharmacol 22:127–140

Begaud B, Martin K, Haramburu F, Moore N (2002) Rates of spontaneous reporting of adverse drug reactions in France. JAMA 288:1588

Behr S, Andersohn F, Garbe E (2010) Risk of intracerebral hemorrhage associated with phenprocoumon exposure: a nested case-control study in a large population-based German database. Pharmacoepidemiol Drug Saf 19:722–730

Behr S, Schill W, Pigeot I (2012) Does additional confounder information alter the estimated risk of bleeding associated with phenprocoumon use-results of a two-phase study. Pharmacoepidemiol Drug Saf 21:535–545

Beiderbeck-Noll AB, Sturkenboom MC, van der Linden PD, Herings RM, Hofman A, Coebergh JW, Leufkens HG, Stricker BH (2003) Verapamil is associated with an increased risk of cancer in the elderly: the Rotterdam study. Eur J Cancer 39:98–105

Benichou C (1994) Imputability of unexpected or toxic drug reactions. The official French method of causality assessment. In: Benichou C (ed) Adverse drug reactions. A practical guide to diagnosis and management. Wiley, Chichester, pp 271–275

Blais L, Ernst P, Boivin JF, Suissa S (1998a) Inhaled corticosteroids and the prevention of readmission to hospital for asthma. Am J Respir Crit Care Med 158:126–132

Blais L, Suissa S, Boivin JF, Ernst P (1998b) First treatment with inhaled corticosteroids and the prevention of admissions to hospital for asthma. Thorax 53:1025–1029

Bloemenkamp KW, Rosendaal FR, Helmerhorst FM, Buller HR, Vandenbroucke JP (1995) Enhancement by factor V Leiden mutation of risk of deep-vein thrombosis associated with oral contraceptives containing a third-generation progestagen. Lancet 346:1593–1596

Bodmer M, Becker C, Meier C, Jick SS, Meier CR (2012) Use of antidiabetic agents and the risk of pancreatic cancer: a case-control analysis. Am J Gastroenterol 107:620–626

Brauchli YB, Jick SS, Meier CR (2011) Statin use and risk of first-time psoriasis diagnosis. J Am Acad Dermatol 65:77–83

Brookhart MA, Wang PS, Solomon DH, Schneeweiss S (2006a) Instrumental variable analysis of secondary pharmacoepidemiologic data. Epidemiology 17:373–374

Brookhart MA, Wang PS, Solomon DH, Schneeweiss S (2006b) Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. Epidemiology 17:268–275

Brookhart MA, Rassen JA, Schneeweiss S (2010) Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiol Drug Saf 19:537–554

Carson JL, Strom BL, Soper KA, West SL, Morse ML (1987) The association of nonsteroidal anti-inflammatory drugs with upper gastrointestinal tract bleeding. Arch Intern Med 147:85–88

Chen H, Zhang SM, Hernan MA, Schwarzschild MA, Willett WC, Colditz GA, Speizer FE, Ascherio A (2003) Nonsteroidal anti-inflammatory drugs and the risk of Parkinson disease. Arch Neurol 60:1059–1064

Clinical Practice Research Datalink (2012) Welcome to CPRD – Clinical Practice Research Datalink. http://www.cprd.com. Accessed 9 Feb 2012

Colditz GA, Hankinson SE, Hunter DJ, Willett WC, Manson JE, Stampfer MJ, Hennekens C, Rosner B, Speizer FE (1995) The use of estrogens and progestins and the risk of breast cancer in postmenopausal women. N Engl J Med 332:1589–1593

Confavreux C, Suissa S, Saddier P, Bourdes V, Vukusic S (2001) Vaccinations and the risk of relapse in multiple sclerosis. Vaccines in Multiple Sclerosis Study Group. N Engl J Med 344:319–326

Corrao G, Zambon A, Faini S, Bagnardi V, Leoni O, Suissa S (2005) Short-acting inhaled beta-2-agonists increased the mortality from chronic obstructive pulmonary disease in observational designs. J Clin Epidemiol 58:92–97

de la Porte M, Reith D, Tilyard M (2002) Impact of safety alerts upon prescribing of cisapride to children in New Zealand. N Z Med J 115:U24

Dietlein G, Schroder-Bernhardi D (2002) Use of the mediplus patient database in healthcare research. Int J Clin Pharmacol Ther 40:130–133

Dietlein G, Schroder-Bernhardi D (2003) Doctors' prescription behaviour regarding dosage recommendations for preparations of kava extracts. Pharmacoepidemiol Drug Saf 12:417–421

Donahue JG, Weiss ST, Livingston JM, Goetsch MA, Greineder DK, Platt R (1997) Inhaled steroids and the risk of hospitalization for asthma. JAMA 277:887–891

Donnan PT, Wang J (2001) The case-crossover and case-time-control designs in pharmacoepidemiology. Pharmacoepidemiol Drug Saf 10:259–262

Downey W, Beck P, McNutt M, Stang M, Osei W, Nichol J (2003) Health databases in Saskatchewan. In: Strom BL (ed) Pharmacoepidemiology. Wiley, Chichester, pp 325–345

DuMouchel W (1999) Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Stat 53:177–202

Ekedahl A, Mansson N (2004) Unclaimed prescriptions after automated prescription transmittals to pharmacies. Pharm World Sci 26:26–31

Eland IA, Belton KJ, van Grootheest AC, Meiners AP, Rawlins MD, Stricker BH (1999) Attitudinal survey of voluntary reporting of adverse drug reactions. Br J Clin Pharmacol 48:623–627

Esposito D, Schone E, Williams T, Liu S, Cybulski K, Stapulonis R, Clusen N (2008) Prevalence of unclaimed prescriptions at military pharmacies. J Manag Care Pharm 14:541–552

Essebag V, Genest J Jr, Suissa S, Pilote L (2003) The nested case-control study in cardiology. Am Heart J 146:581–590

Etienney I, Beaugerie L, Viboud C, Flahault A (2003) Non-steroidal anti-inflammatory drugs as a risk factor for acute diarrhoea: a case crossover study. Gut 52:260–263

Evans SJ, Waller PC, Davis S (2001) Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf 10:483–486

Fagot JP, Mockenhaupt M, Bouwes-Bavinck JN, Naldi L, Viboud C, Roujeau JC (2001) Nevirapine and the risk of Stevens-Johnson syndrome or toxic epidermal necrolysis. AIDS 15:1843–1848

Faich GA (1986) Adverse-drug-reaction monitoring. N Engl J Med 314:1589–1592

Fairfield KM, Hunter DJ, Fuchs CS, Colditz GA, Hankinson SE (2002) Aspirin, other NSAIDs, and ovarian cancer risk (United States). Cancer Causes Control 13:535–542

Farrington CP (1995) Relative incidence estimation from case series for vaccine safety evaluation. Biometrics 51:228–235

Farrington P, Pugh S, Colville A, Flower A, Nash J, Morgan-Capner P, Rush M, Miller E (1995) A new method for active surveillance of adverse events from diphtheria/tetanus/pertussis and measles/mumps/rubella vaccines. Lancet 345:567–569

Feenstra J, Heerdink ER, Grobbee DE, Stricker BH (2002) Association of nonsteroidal anti-inflammatory drugs with first occurrence of heart failure and with relapsing heart failure: the Rotterdam study. Arch Intern Med 162:265–270

Felson DT, Sloutskis D, Anderson JJ, Anthony JM, Kiel DP (1991) Thiazide diuretics and the risk of hip fracture. Results from the Framingham Study. JAMA 265:370–373

Friedman DE, Habel LA, Boles M, McFarland BH (2000) Kaiser Permanente Medical Care Program: Division of Research, Northern California, and Center for Health Research, Northwest Division. In: Strom BL (ed) Pharmacoepidemiology. Wiley, Chichester, pp 263–283

Friedman MA, Woodcock J, Lumpkin MM, Shuren JE, Hass AE, Thompson LJ (1999) The safety of newly approved medicines: do recent market removals mean there is a problem? JAMA 281:1728–1734

Funkhouser EM, Sharp GB (1995) Aspirin and reduced risk of esophageal carcinoma. Cancer 76:1116–1119

Furu K (2008) Establishment of the nationwide Norwegian Prescription Database (NorPD) – new opportunities for research in pharmacoepidemiology in Norway. Norsk Epidemiologi 18: 129–136

Furu K, Wettermark B, Andersen M, Martikainen JE, Almarsdottir AB, Sorensen HT (2010) The Nordic countries as a cohort for pharmacoepidemiological research. Basic Clin Pharmacol Toxicol 106:86–94

Garbe E, LeLorier J, Boivin JF, Suissa S (1997) Inhaled and nasal glucocorticoids and the risks of ocular hypertension or open-angle glaucoma. JAMA 277:722–727

Garbe E, Boivin JF, LeLorier J, Suissa S (1998a) Selection of controls in database case-control studies: glucocorticoids and the risk of glaucoma. J Clin Epidemiol 51:129–135

Garbe E, Suissa S, LeLorier J (1998b) Association of inhaled corticosteroid use with cataract extraction in elderly patients. JAMA 280:539–543

Garbe E, Suling M, Kloss S, Lindemann C, Schmid U (2011) Linkage of mother-baby pairs in the German Pharmacoepidemiological Research Database. Pharmacoepidemiol Drug Saf 20:258–264

Garcia Rodriguez LA, Perez GS (1998) Use of the UK General Practice Research Database for pharmacoepidemiology. Br J Clin Pharmacol 45:419–425

Giovannucci E, Rimm EB, Stampfer MJ, Colditz GA, Ascherio A, Willett WC (1994) Aspirin use and the risk for colorectal cancer and adenoma in male health professionals. Ann Intern Med 121:241–246

Glessner MR, Heller DA (2002) Changes in related drug class utilization after market withdrawal of cisapride. Am J Manag Care 8:243–250

GPRD (2012) The Gold Standard of Healthcare Data. http://www.gprd.com/gprd/goldstandard.asp. Accessed 9 Feb 2012

Greenland S (1996) Confounding and exposure trends in case-crossover and case-time-control design. Epidemiology 7:231–239

Griffin JP (1986) Survey of the spontaneous adverse drug reaction reporting schemes in fifteen countries. Br J Clin Pharmacol 22:83S–100S

Grodstein F, Stampfer MJ, Manson JE, Colditz GA, Willett WC, Rosner B, Speizer FE, Hennekens CH (1996) Postmenopausal estrogen and progestin use and the risk of cardiovascular disease. N Engl J Med 335:453–461

Grodstein F, Martinez ME, Platz EA, Giovannucci E, Colditz GA, Kautzky M, Fuchs C, Stampfer MJ (1998) Postmenopausal hormone use and risk for colorectal cancer and adenoma. Ann Intern Med 128:705–712

Group Health (2012) About us. http://www.ghc.org/about_gh/index.jhtml. Accessed 8 Feb 2012

Haramburu F, Tubert-Bitter P, Begaud B (1992) Trends in spontaneous reporting. Post Mark Surveill 6:129–134

Haramburu F, Begaud B, Moride Y (1997) Temporal trends in spontaneous reporting of unlabelled adverse drug reactions. Br J Clin Pharmacol 44:299–301

Hartnell NR, Wilson JP (2004) Replication of the Weber effect using postmarketing adverse event reports voluntarily submitted to the United States Food and Drug Administration. Pharmacotherapy 24:743–749

Hee KJ, Grodstein F (2003) Regular use of nonsteroidal anti-inflammatory drugs and cognitive function in aging women. Neurology 60:1591–1597

Hennessy S, Strom BL (2000) Nonsedating antihistamines should be preferred over sedating antihistamines in patients who drive. Ann Intern Med 132:405–407

Hernan MA, Robins JM (2006) Instruments for causal inference: an epidemiologist's dream? Epidemiology 17:360–372

Hernan MA, Hohol MJ, Olek MJ, Spiegelman D, Ascherio A (2000) Oral contraceptives and the incidence of multiple sclerosis. Neurology 55:848–854

Hernandez-Avila M, Liang MH, Willett WC, Stampfer MJ, Colditz GA, Rosner B, Chang RW, Hennekens CH, Speizer FE (1990) Exogenous sex hormones and the risk of rheumatoid arthritis. Arthritis Rheum 33:947–953

Hernandez-Diaz S, Hernan MA, Meyer K, Werler MM, Mitchell AA (2003) Case-crossover and case-time-control designs in birth defects epidemiology. Am J Epidemiol 158:385–391

Horwitz RI, Feinstein AR (1981) Improved observational method for studying therapeutic efficacy. Suggestive evidence that lidocaine prophylaxis prevents death in acute myocardial infarction. JAMA 246:2455–2459

Hubbard RB, Smith CJ, Smeeth L, Harrison TW, Tattersfield AE (2002) Inhaled corticosteroids and hip fracture: a population-based case-control study. Am J Respir Crit Care Med 166:1563–1566

Hubbard R, Farrington P, Smith C, Smeeth L, Tattersfield A (2003) Exposure to tricyclic and selective serotonin reuptake inhibitor antidepressants and the risk of hip fracture. Am J Epidemiol 158:77–84

IPCI (2012a) Interdisciplinary Processing of Clinical Information (IPCI) – information. http://www.ipci.nl/Framework/Frames.php?language=UK&subsite=Publications. Accessed 13 Feb 2012

IPCI (2012b) Interdisciplinary Processing of Clinical Information. http://www.ipci.nl. Accessed 13 Feb 2012

Jick H, Jick SS, Gurewich V, Myers MW, Vasilakis C (1995) Risk of idiopathic cardiovascular death and nonfatal venous thromboembolism in women using oral contraceptives with differing progestagen components. Lancet 346:1589–1593

Jobski K, Behr S, Garbe E (2011) Drug interactions with phenprocoumon and the risk of serious haemorrhage: a nested case-control study in a large population-based German database. Eur J Clin Pharmacol 67:941–951

Kaiser Permanente (2012a) Who we are. http://www.kaiserpermanentejobs.org/northern-california.aspx. Accessed 8 Feb 2012

Kaiser Permanente (2012b) Press releases: Northwest. http://xnet.kp.org/newscenter/pressreleases/nw/2011/062711gateway.html. Accessed 8 Feb 2012

Kaufman DW, Kelly JP, Jurgelon JM, Anderson T, Issaragrisil S, Wiholm BE, Young NS, Leaverton P, Levy M, Shapiro S (1996) Drugs in the aetiology of agranulocytosis and aplastic anaemia. Eur J Haematol Suppl 60:23–30

KELA – The Social Insurance Institution of Finland (2012) Statistics on reimbursements for medical expenses. http://www.kela.fi/in/internet/english.nsf/NET/151008123439AS?OpenDocument. Accessed 14 Feb 2012

Kelly JP, Rosenberg L, Kaufman DW, Shapiro S (1990) Reliability of personal interview data in a hospital-based case-control study. Am J Epidemiol 131:79–90

Kennedy J, Tuleu I, Mackay K (2008) Unfilled prescriptions of medicare beneficiaries: prevalence, reasons, and types of medicines prescribed. J Manag Care Pharm 14:553–560

Ki M, Park T, Yi SG, Oh JK, Choi B (2003) Risk analysis of aseptic meningitis after measles-mumps-rubella vaccination in Korean children by using a case-crossover design. Am J Epidemiol 157:158–165

Kiel DP, Felson DT, Anderson JJ, Wilson PW, Moskowitz MA (1987) Hip fracture and the use of estrogens in postmenopausal women. The Framingham Study. N Engl J Med 317:1169–1174

Kirking DM, Lee JA, Ellis JJ, Briesacher B, McKercher PL (2006) Patient-reported underuse of prescription medications: a comparison of nine surveys. Med Care Res Rev 63:427–446

Kjaer D, Horvath-Puho E, Christensen J, Vestergaard M, Czeizel AE, Sorensen HT, Olsen J (2007) Use of phenytoin, phenobarbital, or diazepam during pregnancy and risk of congenital abnormalities: a case-time-control study. Pharmacoepidemiol Drug Saf 16:181–188

Kromann-Andersen H, Pedersen A (1988) Reported adverse reactions to and consumption of nonsteroidal anti-inflammatory drugs in Denmark over a 17-year period. Dan Med Bull 35:187–192

Lando JF, Heck KE, Brett KM (1999) Hormone replacement therapy and breast cancer risk in a nationally representative cohort. Am J Prev Med 17:176–180

Lee D, Bergman U (2000) Studies of drug utilization. In: Strom BL (ed) Pharmacoepidemiology. Wiley, Chichester, pp 463–481

Lenz W (1987) The Thalidomide hypothesis: how it was found and tested. In: Kewitz H, Roots I, Voigt K (eds) Epidemiological concepts in clinical pharmacology. Springer, Heidelberg, pp 3–10

Leone R, Venegoni M, Motola D, Moretti U, Piazzetta V, Cocci A, Resi D, Mozzo F, Velo G, Burzilleri L, Montanaro N, Conforti A (2003) Adverse drug reactions related to the use of fluoroquinolone antimicrobials: an analysis of spontaneous reports and fluoroquinolone consumption data from three italian regions. Drug Saf 26:109–120

Levesque LE, Hanley JA, Kezouh A, Suissa S (2010) Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. BMJ 340:b5087

Lewis MA, Heinemann LA, MacRae KD, Bruppacher R, Spitzer WO (1996) The increased risk of venous thromboembolism and the use of third generation progestagens: role of bias in observational research. The Transnational Research Group on Oral Contraceptives and the Health of Young Women. Contraception 54:5–13

Lindquist M, Edwards IR (1993) Adverse drug reaction reporting in Europe: some problems of comparisons. Int J Risk Saf Med 4:35–46

MacDonald TM, Morant SV, Goldstein JL, Burke TA, Pettitt D (2003) Channelling bias and the incidence of gastrointestinal haemorrhage in users of meloxicam, coxibs, and older, non-specific non-steroidal anti-inflammatory drugs. Gut 52:1265–1270

Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH (2006) Instrumental variables: application and limitations. Epidemiology 17:260–267

McLeod PJ, Huang AR, Tamblyn RM, Gayton DC (1997) Defining inappropriate practices in prescribing for elderly people: a national consensus panel. CMAJ 156:385–391

Meier C, Brauchli YB, Jick SS, Kraenzlin ME, Meier CR (2010) Use of depot medroxyprogesterone acetate and fracture risk. J Clin Endocrinol Metab 95:4909–4916

MHRA Centre NorthWest (2009) Yellow Card. http://www.yccnorthwest.nhs.uk/reporting.aspx. Accessed 10 May 2012

Miettinen OS (1983) The need for randomization in the study of intended effects. Stat Med 2: 267–271

Mojtabai R, Zivin JG (2003) Effectiveness and cost-effectiveness of four treatment modalities for substance disorders: a propensity score analysis. Health Serv Res 38:233–259

Moore N, Hall G, Sturkenboom M, Mann R, Lagnaoui R, Begaud B (2003) Biases affecting the proportional reporting ratio (PPR) in spontaneous reports pharmacovigilance databases: the example of sertindole. Pharmacoepidemiol Drug Saf 12:271–281

Moride Y, Abenhaim L (1994) Evidence of the depletion of susceptibles effect in non-experimental pharmacoepidemiologic research. J Clin Epidemiol 47:731–737

Neutel CI, Perry S, Maxwell C (2002) Medication use and risk of falls. Pharmacoepidemiol Drug Saf 11:97–104

Nicholas JM, Grieve AP, Gulliford MC (2012) Within-person study designs had lower precision and greater susceptibility to bias because of trends in exposure than cohort and nested case-control designs. J Clin Epidemiol 65:384–393

Opatrny L, Delaney JA, Suissa S (2008) Gastro-intestinal haemorrhage risks of selective serotonin receptor antagonist therapy: a new look. Br J Clin Pharmacol 66:76–81

Park-Wyllie LY, Mamdani MM, Li P, Gill SS, Laupacis A, Juurlink DN (2009) Cholinesterase inhibitors and hospitalization for bradycardia: a population-based study. PLoS Med 6:e1000157

Pedianet Project (2012) Pedianet database, general design. http://www.pedianet.it/en/pedianet-database/general-design/;213. Accessed 13 Feb 2012

Perez E, Schroder-Bernhardi D, Dietlein G (2002) Treatment behavior of doctors regarding Helicobacter pylori infections. Int J Clin Pharmacol Ther 40:126–129

Perkins SM, Tu W, Underhill MG, Zhou XH, Murray MD (2000) The use of propensity scores in pharmacoepidemiologic research. Pharmacoepidemiol Drug Saf 9:93–101

PHARMO Institute for Drug Outcomes Research (2012) Company history. http://www.pharmo.nl/common/page/history. Accessed 13 Feb 2012

Pierfitte C, Begaud B, Lagnaoui R, Moore ND (1999) Is reporting rate a good predictor of risks associated with drugs? Br J Clin Pharmacol 47:329–331

Pierfitte C, Royer RJ, Moore N, Begaud B (2000) The link between sunshine and phototoxicity of sparfloxacin. Br J Clin Pharmacol 49:609–612

Pigeot I, Ahrens W (2008) Establishment of a pharmacoepidemiological database in Germany: methodological potential, scientific value and practical limitations. Pharmacoepidemiol Drug Saf 17:215–223

Porta M, Hartzema AG, Tilson HH (1997) The contribution of epidemiology to the study of drug effects. In: Hartzema AG, Porta M, Tilson HH (eds) Pharmacoepidemiology. The fundamentals. Harvey Whitney, Cincinnati, pp 1–28

QResearch (2012a) QRESEARCH specialises in research & analyses using primary care electronic health data. http://www.qresearch.org. Accessed 12 March 2012

QResearch (2012b) What is QRESEARCH? http://www.qresearch.org/SitePages/What%20Is%20QResearch.aspx. Accessed 13 Feb 2012

Raiford DS, Perez GS, Garcia Rodriguez LA (1996) Positive predictive value of ICD-9 codes in the identification of cases of complicated peptic ulcer disease in the Saskatchewan hospital automated database. Epidemiology 7:101–104

RAMQ (2010) Historical background; yesterday and today. http://www.ramq.gouv.qc.ca/en/regie/historique/hier_auj.shtml. Accessed 8 Feb 2012

Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S (2009a) Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. J Clin Epidemiol 62:1226–1232

Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S (2009b) Instrumental variables II: instrumental variable application-in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance. J Clin Epidemiol 62:1233–1241

Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S (2011) Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. Am J Epidemiol 173:1404–1413

Rawlins MD, Breckenridge AM, Wood SM (1989) National adverse drug reaction reporting – a silver jubilee. Advers Drug React Bull 138:516–519

Ray WA (2003) Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol 158:915–920

Ray WA, Griffin MR (1989) Use of Medicaid data for pharmacoepidemiology. Am J Epidemiol 129:837–849

Ray WA, Griffin MR, Downey W, Melton LJ III (1989) Long-term use of thiazide diuretics and risk of hip fracture. Lancet 1:687–690

Risselada R, Straatman H, van Kooten F, Dippel DW, van der Lugt A, Niessen WJ, Firouzian A, Herings RM, Sturkenboom MC (2011) Platelet aggregation inhibitors, vitamin K antagonists and risk of subarachnoid hemorrhage. J Thromb Haemost 9:517–523

Rothman KJ, Lanes S, Sacks ST (2004) The reporting odds ratio and its advantages over the proportional reporting ratio. Pharmacoepidemiol Drug Saf 13:519–523

Samet JM (2003) Measuring the effectiveness of inhaled corticosteroids for COPD is not easy! Am J Respir Crit Care Med 168:1–2

Saskatchewan Ministry of Health (2010) Health services databases: information document. http://www.health.gov.sk.ca/health-databases-info-doc. Accessed 24 March 2012

Saunders KW, Davis RL, Stergachis A (2000) Group Health Cooperative of Puget Sound. In: Strom BL (ed) Pharmacoepidemiology. Wiley, Chichester, pp 247–262

Schade R, Andersohn F, Suissa S, Haverkamp W, Garbe E (2007) Dopamine agonists and the risk of cardiac-valve regurgitation. N Engl J Med 356:29–38

Schneeweiss S (2006) Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiol Drug Saf 15:291–303

Schneeweiss S, Avorn J (2005) A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 58:323–337

Schneeweiss S, Rassen J (2011) Re: confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. Pharmacoepidemiol Drug Saf 20:1110–1111

Schneeweiss S, Setoguchi S, Brookhart A, Dormuth C, Wang PS (2007) Risk of death associated with the use of conventional versus atypical antipsychotic drugs among elderly patients. CMAJ 176:627–632

Schneeweiss S, Seeger JD, Landon J, Walker AM (2008) Aprotinin during coronary-artery bypass grafting and risk of death. N Engl J Med 358:771–783

Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA (2009) High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology 20:512–522

Schneider MF, Gange SJ, Margolick JB, Detels R, Chmiel JS, Rinaldo C, Armenian HK (2005) Application of case-crossover and case-time-control study designs in analyses of time-varying predictors of T-cell homeostasis failure. Ann Epidemiol 15:137–144

Schoofs MW, van der Klift M, Hofman A, de Laet CE, Herings RM, Stijnen T, Pols HA, Stricker BH (2003) Thiazide diuretics and the risk for hip fracture. Ann Intern Med 139:476–482

Schroder D, Weiser M, Klein P (2003) Efficacy of a homeopathic Crataegus preparation compared with usual therapy for mild (NYHA II) cardiac insufficiency: results of an observational cohort study. Eur J Heart Fail 5:319–326

Schroder-Bernhardi D, Dietlein G (2002) Lipid-lowering therapy: do hospitals influence the prescribing behavior of general practitioners? Int J Clin Pharmacol Ther 40:317–321

Seeger JD, Walker AM, Williams PL, Saperia GM, Sacks FM (2003) A propensity score-matched cohort study of the effect of statins, mainly fluvastatin, on the occurrence of acute myocardial infarction. Am J Cardiol 92:1447–1451

Serradell J, Bjornson DC, Hartzema AG (1991) Drug utilization studies: sources and methods. In: Hartzema AG, Porta MS, Tilson HH (eds) Pharmacoepidemiology: an introduction. Harvey Whitney, Cincinnati, pp 101–119

Shapiro S (1989) The role of automated record linkage in the postmarketing surveillance of drug safety: a critique. Clin Pharmacol Ther 46:371–386

Sin DD, Man SF (2002) Low-dose inhaled corticosteroid therapy and risk of emergency department visits for asthma. Arch Intern Med 162:1591–1595

Sin DD, Tu JV (2001a) Inhaled corticosteroid therapy reduces the risk of rehospitalization and all-cause mortality in elderly asthmatics. Eur Respir J 17:380–385

Sin DD, Tu JV (2001b) Inhaled corticosteroids and the risk of mortality and readmission in elderly patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 164:580–584

Slone D, Shapiro S, Miettinen OS, Finkle WD, Stolley PD (1979) Drug evaluation after marketing. Ann Intern Med 90:257–261

Spitzer WO, Suissa S, Ernst P, Horwitz RI, Habbick B, Cockcroft D, Boivin JF, McNutt M, Buist AS, Rebuck AS (1992) The use of beta-agonists and the risk of death and near death from asthma. N Engl J Med 326:501–506

Spitzer WO, Lewis MA, Heinemann LA, Thorogood M, MacRae KD (1996) Third generation oral contraceptives and risk of venous thromboembolic disorders: an international case-control study. Transnational Research Group on Oral Contraceptives and the Health of Young Women. BMJ 312:83–88

State of New Jersey (2012) Department of Human Services, Division of Medical Assistance and Human Services: monthly enrollment reports NJ FamilyCare/Medicaid enrollment statistics; January 2012. http://www.nj.gov/humanservices/dmahs/news/reports/enrollment_2012_01.pdf. Accessed 13 Feb 2012

Stricker BH, Tijssen JG (1992) Serum sickness-like reactions to cefaclor. J Clin Epidemiol 45:1177–1184

Strom BL, Carson JL (1990) Use of automated databases for pharmacoepidemiology research. Epidemiol Rev 12:87–107

Strom BL, Miettinen OS, Melmon KL (1983) Postmarketing studies of drug efficacy: when must they be randomized? Clin Pharmacol Ther 34:1–7

Strom BL, Miettinen OS, Melmon KL (1984) Post-marketing studies of drug efficacy: how? Am J Med 77:703–708

Strom BL, Carson JL, Halpern AC, Schinnar R, Snyder ES, Stolley PD, Shaw M, Tilson HH, Joseph M, Dai WS (1991) Using a claims database to investigate drug-induced Stevens-Johnson syndrome. Stat Med 10:565–576

Strom BL, Kimmel SE, Hennessy S (2012) Pharmacoepidemiology. Wiley, Chichester

Sturkenboom M (2007) Other databases in Europe for the analytic evaluation of drug effects. In: Mann RD, Andrews EB (eds) Pharmacovigilance. Wiley, Chichester/Hoboken, pp 362–372

Sturkenboom MC, Middelbeek A, de Jong-van den Berg LT, van den Berg PB, Stricker BH, Wesseling H (1995) Vulvo-vaginal candidiasis associated with acitretin. J Clin Epidemiol 48:991–997

Suissa S (1995) The case-time-control design. Epidemiology 6:248–253

Suissa S (1998) The case-time-control design: further assumptions and conditions [comment]. Epidemiology 9:441–445

Suissa S (2000) Novel approaches to pharmacoepidemiology study design and statistical analysis. In: Strom BL (ed) Pharmacoepidemiology. Wiley, Chichester, pp 785–805

Suissa S (2003) Effectiveness of inhaled corticosteroids in COPD: immortal time bias in observational studies. Am J Respir Crit Care Med 168:49–53

Suissa S (2004) Inhaled steroids and mortality in COPD: bias from unaccounted immortal time. Eur Respir J 23:391–395

Suissa S (2007) Immortal time bias in observational studies of drug effects. Pharmacoepidemiol Drug Saf 16:241–249

Suissa S (2008) Immortal time bias in pharmaco-epidemiology. Am J Epidemiol 167:492–499

Suissa S, Garbe E (2007) Primer: administrative health databases in observational studies of drug effects–advantages and disadvantages. Nat Clin Pract Rheumatol 3:725–732

Suissa S, Blais L, Spitzer WO, Cusson J, Lewis M, Heinemann L (1997) First-time use of newer oral contraceptives and the risk of venous thromboembolism. Contraception 56:141–146

Suissa S, Ernst P, Benayoun B, Baltzan M, Cai B (2000a) Low-dose inhaled corticosteroids and the prevention of death from asthma. N Engl J Med 343:332–336

Suissa S, Spitzer WO, Rainville B, Cusson J, Lewis M, Heinemann L (2000b) Recurrent use of newer oral contraceptives and the risk of venous thromboembolism. Hum Reprod 15:817–821

Suissa S, Ernst P, Kezouh A (2002) Regular use of inhaled corticosteroids and the long term prevention of hospitalisation for asthma. Thorax 57:880–884

Suissa S, Baltzan M, Kremer R, Ernst P (2004) Inhaled and nasal corticosteroid use and the risk of fracture. Am J Respir Crit Care Med 169:83–88

Szarfman A, Machado SG, O'Neill RT (2002) Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. Drug Saf 25:381–392

Tamblyn R, Lavoie G, Petrella L, Monette J (1995) The use of prescription claims databases in pharmacoepidemiological research: the accuracy and comprehensiveness of the prescription claims database in Quebec. J Clin Epidemiol 48:999–1009

Tamblyn R, McLeod P, Hanley JA, Girard N, Hurley J (2003) Physician and practice characteristics associated with the early utilization of new prescription drugs. Med Care 41:895–908

Taylor B, Miller E, Farrington CP, Petropoulos MC, Favot-Mayaud I, Li J, Waight PA (1999) Autism and measles, mumps, and rubella vaccine: no epidemiological evidence for a causal association. Lancet 353:2026–2029

Tennessee Government (2012) What's new with TennCare. http://www.tn.gov/tenncare/. Accessed 8 Feb 2012

Tennis P, Bombardier C, Malcolm E, Downey W (1993) Validity of rheumatoid arthritis diagnoses listed in the Saskatchewan Hospital Separations Database. J Clin Epidemiol 46:675–683

The Danish Medicines Agency (2010) About the Register of Medicinal Product Statistics. http://laegemiddelstyrelsen.dk/en/topics/statistics,-prices-and-reimbursement/statistics-and-analyses/about-the-register-of-medicinal-product—tatistics-. Accessed 14 Feb 2012

The Health Improvement Network (2012) Welcome to The Health Improvement Network website. http://www.thin-uk.com. Accessed 12 March 2012

The National Board of Health and Welfare (Socialstyrelsen) (2012a) Statistics on health, use of health services, social conditions and social services. http://192.137.163.40/epcfs/index.asp?kod=engelska. Accessed 13 Feb 2012

The National Board of Health and Welfare (Socialstyrelsen) (2012b) The National Patient Register. http://www.socialstyrelsen.se/register/halsodataregister/patientregistret/inenglish. Accessed 14 Feb 2012

Tsong Y (1995) Comparing reporting rates of adverse events between drugs with adjustment for year of marketing and secular trends in total reporting. J Biopharm Stat 5:95–114

UCL Research Department of Primary Care and Population Health (2011) The THIN Database. http://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database. Accessed 10 Feb 2012

Uppsala Monitoring Centre (2011) Vigibase – now over 7 million ADR reports. http://www.umc-products.com/DynPage.aspx?id=75618&news=10614. Accessed 10 May 2012

US Food and Drug Administration (2010) Adverse Event Reporting System (AERS) statistics. http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070093.htm. Accessed 15 July 2011

van den Brandt PA, Petri H, Dorant E, Goldbohm RA, van de Crommert S (1991) Comparison of questionnaire information and pharmacy data on drug use. Pharm Weekbl Sci 13:91–96

van der Kroef C (1979) Reactions to triazolam. Lancet 2:526

van Puijenbroek EP, Bate A, Leufkens HG, Lindquist M, Orre R, Egberts AC (2002) A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf 11:3–10 van Staa TP, Abenhaim L, Leufkens H (1994) A study of the effects of exposure misclassification due to the time-window design in pharmacoepidemiologic studies. J Clin Epidemiol 47:183–189

Venning GR (1983) Identification of adverse reactions to new drugs. III: alerting processes and early warning systems. Br Med J (Clin Res Ed) 286:458–460

Viscoli CM, Brass LM, Kernan WN, Sarrel PM, Suissa S, Horwitz RI (2001) A clinical trial of estrogen-replacement therapy after ischemic stroke. N Engl J Med 345:1243–1249

Vlahovic-Palcevski V, Wettermark B, Bergman U (2002) Quality of non-steroidal anti-inflammatory drug prescribing in Croatia (Rijeka) and Sweden (Stockholm). Eur J Clin Pharmacol 58:209–214

Waller P, van Puijenbroek E, Egberts A, Evans S (2004) The reporting odds ratio versus the proportional reporting ratio: "deuce." Pharmacoepidemiol Drug Saf 13:525–526

Wang J, Donnan PT (2001) Propensity score methods in drug safety studies: practice, strengths and limitations. Pharmacoepidemiol Drug Saf 10:341–344

Wang J, Donnan PT, Steinke D, MacDonald TM (2001) The multiple propensity score for analysis of dose-response relationships in drug safety studies. Pharmacoepidemiol Drug Saf 10:105–111

Wang PS, Schneeweiss S, Avorn J, Fischer MA, Mogun H, Solomon DH, Brookhart MA (2005) Risk of death in elderly users of conventional vs. atypical antipsychotic medications. N Engl J Med 353:2335–2341

Wang PS, Schneeweiss S, Setoguchi S, Patrick A, Avorn J, Mogun H, Choudhry NK, Brookhart MA (2007) Ventricular arrhythmias and cerebrovascular events in the elderly using conventional and atypical antipsychotic medications. J Clin Psychopharmacol 27:707–710

Wang S, Linkletter C, Dore D, Mor V, Buka S, Maclure M (2012) Age, antipsychotics, and the risk of ischemic stroke in the Veterans Health Administration. Stroke 43:28–31

Weatherby LB, Walker AM, Fife D, Vervaet P, Klausner MA (2001) Contraindicated medications dispensed with cisapride: temporal trends in relation to the sending of "Dear Doctor" letters. Pharmacoepidemiol Drug Saf 10:211–218

Weber JCP (1984) Epidemiology of adverse reactions to nonsteroidal antiinflammatory drugs. In: Rainsford KD, Velo GP (eds) Advances in inflammation research. Raven Press, New York, pp 1–6

Weintraub JM, Taylor A, Jacques P, Willett WC, Rosner B, Colditz GA, Chylack LT, Hankinson SE (2002) Postmenopausal hormone use and lens opacities. Ophthalmic Epidemiol 9:179–190

Weldeselassie YG, Whitaker HJ, Farrington CP (2011) Use of the self-controlled case-series method in vaccine safety studies: review and recommendations for best practice. Epidemiol Infect 139:1805–1817

West SL (1997) A comparison of data sources for drug exposure ascertainment in pharmacoepidemiologic studies with emphasis on self-reported information. Pharmacoepidemiol Drug Saf 6:215–218

West SL, Strom BL, Freundlich B, Normand E, Koch G, Savitz DA (1994) Completeness of prescription recording in outpatient medical records from a health maintenance organization. J Clin Epidemiol 47:165–171

West SL, Savitz DA, Koch G, Strom BL, Guess HA, Hartzema A (1995) Recall accuracy for prescription medications: self-report compared with database information. Am J Epidemiol 142:1103–1112

Whitaker HJ, Farrington CP, Spiessens B, Musonda P (2006) Tutorial in biostatistics: the self-controlled case series method. Stat Med 25:1768–1797

Whitaker HJ, Hocine MN, Farrington CP (2009) The methodology of self-controlled case series studies. Stat Methods Med Res 18:7–26

Wiholm B, Olsson S, Moore N, Waller P (2000) Spontaneous reporting systems outside the US. In: Strom BL (ed) Pharmacoepidemiology. Wiley, Chichester, pp 175–192

World Health Organization (1977) The selection of essential drugs. Report of a WHO expert committee. Ser no 615. World Health Organization, Geneva

Worzala K, Hiller R, Sperduto RD, Mutalik K, Murabito JM, Moskowitz M, D'Agostino RB, Wilson PW (2001) Postmenopausal estrogen use, type of menopause, and lens opacities: the Framingham studies. Arch Intern Med 161:1448–1454

Wu K, Willett WC, Fuchs CS, Colditz GA, Giovannucci EL (2002) Calcium intake and risk of colon cancer in women and men. J Natl Cancer Inst 94:437–446

Wysowski DK, Pitts M, Beitz J (2001) An analysis of reports of depression and suicide in patients treated with isotretinoin. J Am Acad Dermatol 45:515–519

Young-Xu Y, Chan KA, Liao JK, Ravid S, Blatt CM (2003) Long-term statin use and psychological well-being. J Am Coll Cardiol 42:690–697

Zambon A, Polo FH, Contiero P, Corrao G (2009) Effect of macrolide and fluoroquinolone antibacterials on the risk of ventricular arrhythmia and cardiac arrest: an observational study in Italy using case-control, case-crossover and case-time-control designs. Drug Saf 32:159–167

Physical Activity Epidemiology 49

Daniela Schmid and Michael F. Leitzmann

Contents

D. Schmid (✉) • M.F. Leitzmann
Department of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_67,

49.1 Introduction

The earliest reports of exercise as a recognized means of health promotion go back to ancient China approximately 2500 BC, where Hua T'o, a Chinese surgeon, encouraged exercises based on the movement of animals. Medical practice in ancient Greece incorporated exercise for the protection and rehabilitation of health through the physicians Hippocrates (460–370 BC) and Galen (129–199 BC). From the sixteenth through the nineteenth century, physical activity was introduced as an important part of hygiene in Western medicine largely by physicians and pharmacologists (MacAuley 1994; U.S. Department of Health and Human Services 1998). The beginnings of modern physical activity epidemiology as a research field can be traced back to the late 1940s, when the British epidemiologist Jeremy Morris uncovered the link between physical exercise and the prevention of heart attacks (Paffenbarger et al. 2001). Finally, the seminal publication of the report of the U.S. Surgeon General, Physical Activity and Health released in 1996, provided public health guidelines and policy and posted important research questions for the future (U.S. Department of Health and Human Services 1996).

Physical activity epidemiology is the study of the distribution and the determinants of physical activity behaviors (Caspersen 1989). It investigates the association of physical activity with disease and other health outcomes and seeks to identify biological, psychosocial, genetic, and environmental factors that influence physical activity. The knowledge obtained from physical activity epidemiology is applied to intervention programs for the prevention and therapy of diseases and for the promotion of health, including population surveillance and evaluation of exercise programs. A major underlying issue in physical activity epidemiology is the development and application of reliable and valid measures of physical activity.

To date, a vast accumulation of epidemiological studies has revealed that moderate exercise prevents chronic illnesses such as cardiovascular disease (CVD), type 2 diabetes, and certain forms of cancer (U.S. Department of Health and Human Services 1998) and that it reduces all-cause mortality (Kesaniemi et al. 2001). Although the benefits of regular, moderate-intensity activity for these and other health outcomes have been widely accepted, precisely quantifying the levels of physical activity is a difficult task. Measurement of physical activity not merely aims to attain accurate assessments of energy expenditure (EE) but also strives to capture dimensions of physical activity including type, duration, frequency, and intensity. The choice of an appropriate assessment method is of major concern as inadequate or crude measurements of physical activity are likely to result in misleading data and may have inappropriate consequences. Two types of methods for measurement of physical activity can be distinguished: (1) objective methods including activity monitors such as pedometers and accelerometers, heart rate monitoring, doubly labeled water (DLW), indirect calorimetry, and direct observation and (2) subjective methods including questionnaires, recalls, diaries, and logs (Vanhees et al. 2005).

The aim of the current chapter is to focus on methods for the measurement of physical activity rather than discuss health outcomes in relation to physical activity, which are described elsewhere (Dishman et al. 2004).

We will begin this chapter with a description of conceptual definitions and the multidimensional aspects of physical activity. Next, we will present domains of physical activity research and physical activity recommendations. Methods and monitoring devices for measuring physical activity that are commonly used in epidemiological studies will be reviewed. We will further address study designs that are commonly used in physical activity research. Another issue will be methodological considerations of studies on physical activity and disease. Finally, we will review areas of potential future development in physical activity epidemiology.

49.2 Definition of Physical Activity

The terms *physical activity*, *physical exercise*, and *physical fitness* have been defined in different ways and have frequently been used interchangeably in physical activity epidemiology. However, a standardized terminology and a clear distinction between these variables are important for the measurement of their associations and interactions with different health outcomes. Physical activity is considered any bodily movement produced by the contraction of skeletal muscle that increases energy expenditure (Caspersen et al. 1985). Physical exercise solely represents a subtype of physical activity undertaken in leisure time that intends to maintain or improve physical fitness (Pate et al. 1995). Physical fitness is distinct from physical activity in that it does not represent a behavior but rather portrays the capacity to attain a certain performance standard or trait (Caspersen et al. 1985; Livingstone et al. 2003). Recently, sedentary behavior has emerged as an additional potential health determinant and has thus become a distinct dimension in physical activity epidemiology.

49.2.1 Physical Activity

Physical activity comprises different types and components including frequency, intensity, and duration which are illustrated in Fig. 49.1. In contrast to physical fitness which describes a physiological construct, physical activity corresponds to a health-related behavior with intra-individual day-to-day variability. Common units of time used to assess physical activity are per day or per week.

Whether the term physical activity comprises all activities or comprises only defined activities in specific domains (occupational, household, transportation, or recreational) depends on the purpose of the individual study. Occupational activity is defined as any work-related activity such as walking, lifting, carpentry work, hauling, pushing, shoveling, and packing boxes normally within the time frame of an 8-hour workday. In contrast, housework, yard work, physically active child care, and

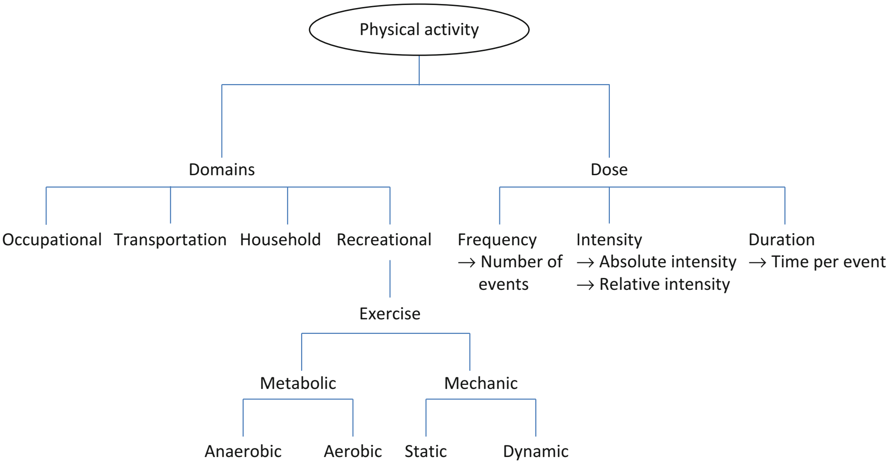


Fig. 49.1 Components of physical activity

chores are covered by the term household activity. Transportation activity includes walking or bicycling for the purposes of going somewhere. Finally, recreational physical activity can be subcategorized into exercise training, competitive sports, and recreation that are undertaken during leisure or discretionary time by personal choice (Courneya et al. 2000).

49.2.2 Physical Exercise

Physical exercise (exercise training) is defined as "a subset of physical activity that is planned, structured, and repetitive and has as a final or an intermediate objective the improvement or maintenance of physical fitness" (Caspersen et al. 1985). It describes a form of leisure-time physical activity involving repetitive bodily movement that is performed to improve or maintain physical fitness. Physical exercise can be categorized according to its mechanical (static or dynamic) or metabolic (aerobic or anaerobic) characteristics. Depending on the muscle contraction that makes the limbs move, the term mechanical differentiates between (1) isometric (same length) or static exercise if there is no movement of the limb and (2) isotonic (same tension) or dynamic exercise if there is movement of the limb (Haskell and Kiernan 2000).

The classification pertaining to the metabolic characteristics of physical exercise encompasses the availability of oxygen (aerobic) or the non-availability of oxygen (anaerobic) for muscle contraction depending on the intensity of exercise performed. Aerobic exercise involves large muscle groups in dynamic activities leading to increases in heart rate and energy expenditure (Haskell and Kiernan 2000). Examples of aerobic exercise include walking, hiking, skating, jogging, cycling, rowing, stair climbing, dancing, and swimming. Anaerobic exercise is performed at a very high-intensity level, and thus, energy is provided by glycolysis and stored phosphocreatine (Howley 2001). Resistance training such as weight lifting is anaerobic exercise that is performed to increase muscular size, strength, and endurance (Hu 2008). Several activities may involve both static and dynamic contractions as well as aerobic and anaerobic metabolism (U.S. Department of Health and Human Services 1998).

Sometimes classification of an activity as exercise may be unclear. For example, using a bicycle solely as transportation to work is not exercise, but bicycling with the idea of reducing body weight is exercise. Hence, it has been suggested that the term *exercise training* is used when activity is performed for the exclusive purpose of enhancing physical fitness (Haskell and Kiernan 2000).

49.2.3 Physical Fitness

Physical fitness is defined as "a set of attributes that people have or achieve that relates to their ability to perform physical activity" and is therefore often used as a surrogate marker for physical activity (Caspersen et al. 1985). Indeed, physical fitness has been shown to be more strongly related to various health outcomes than physical activity (Blair et al. 2001). Its primary modifiable determinant is habitual physical activity, and fitness can thus represent a valuable surrogate marker for physical activity. Several other factors may influence cardiorespiratory fitness including sex, age, health status, and genetic components (Sui et al. 2007).

Observations from the HERITAGE Family study revealed that the heritability of maximal oxygen uptake ($VO_{2\,max}$) among sedentary adults can reach 50%, although that value is potentially inflated by non-genetic factors (Bouchard et al. 1998). Furthermore, maternal influence has been shown to account for nearly 30% of familial transmission. Bouchard et al. (1999) investigated whether individual differences in responses of $VO_{2\,max}$ to a standardized training program are characterized by familial aggregation (Bouchard et al. 1999). In their study, the maximal heritability estimate of the $VO_{2\,max}$ response to exercise training reached 47%, with maternal heritability reaching 28%.

As the attributes related to physical fitness differ in their importance for athletic performance versus their relevance to health, physical fitness has been classified into performance-related fitness and health-related fitness (Pate et al. 1995) (Fig. 49.2). Performance-related fitness is usually linked to attributes such as power, balance, and reaction time. Health-related fitness includes cardiorespiratory fitness, also termed aerobic fitness or endurance, and body composition, muscular fitness, and flexibility (U.S. Department of Health and Human Services 1998). Cardiorespiratory fitness represents the major component of physical fitness, and quantitative data on the relationships between physical fitness and health are mostly based on cardiorespiratory fitness and body composition (Garber et al. 2011). Cardiorespiratory fitness refers to the ability of the circulatory and respiratory systems to supply oxygen to the working muscles during heavy dynamic exercise. During physical exertion, oxygen uptake (VO_2) increases directly proportional to the rate of work. The $VO_{2\,max}$ that can be utilized by the body during extensive

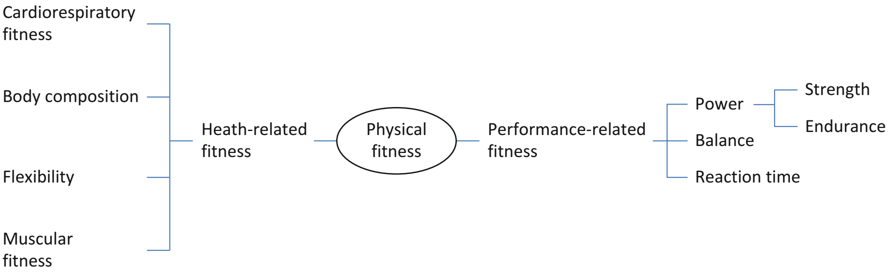


Fig. 49.2 Components of physical fitness

physical effort represents the most accurate measure of cardiorespiratory fitness (U.S. Department of Health and Human Services 1998). Physical fitness assessment has the disadvantage that it is not feasible in large studies compared to physical activity measures by questionnaires.

49.2.4 Sedentary Behavior

In the past decade, interest in sedentary behavior as an independent health risk factor has emerged. According to the definition by Patel et al. (2008), sedentary behavior refers to activities with low levels of energy expenditure in the range of 1.0 to 1.5 metabolic equivalent tasks (METs) (Patel et al. 2008). That definition corresponds to activities such as sleeping, sitting, lying down, watching TV, and other forms of screen-based entertainment (Patel et al. 2008). Among sedentary behaviors, TV viewing has been the most commonly studied.

It is apparent that sedentary behavior is not merely the absence of physical activity (Tudor-Locke and Myers 2001; Healy et al. 2008). In epidemiological studies, the correlation between sedentary behavior and physical activity has tended to be minimal (Hu et al. 2003). Moreover, time spent in sedentary behaviors was found to have correlates distinct from those associated with physical activity (Leatherdale and Wong 2008). Moreover, sedentary behavior may easily coexist with physical activity (Owen et al. 2010). However, misconceptions about the definition and the measurement of sedentary behavior exist. Thus, more research on the assessment of sedentary behavior and its potential health outcomes is needed.

49.3 Dimensions of Physical Activity

49.3.1 Energy Expenditure of Physical Activity

Energy expenditure reflects the energy cost or intensity due to physical activity and it comprises a complex series of biochemical processes and physiological adaptations that result in the transfer of metabolic energy to perform skeletal

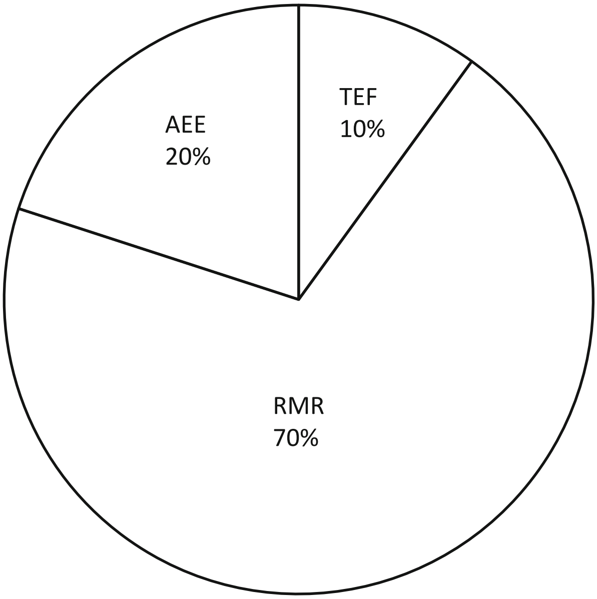


Fig. 49.3 Distribution of major components of total energy expenditure; *AEE* activity energy expenditure, *TEF* thermic effect of food, *RMR* resting metabolic rate

muscle contraction (Lamonte and Ainsworth 2001; Lee et al. 2009). Several factors may influence energy expenditure, such as age, body size, body weight, and fitness level (Lamonte and Ainsworth 2001). Typically, total energy expenditure (TEE) is divided into three components: (1) resting metabolic rate (RMR) as the main component (approximately 70% of TEE), (2) thermic effect of food (TEF, approximately 10% of TEE), and (3) activity energy expenditure due to physical activity or muscular activity (AEE, typically 20% of TEE), which represents the most variable component of TEE (Fig. 49.3).

The rate at which the body uses energy is expressed as the metabolic rate. It can be estimated by indirect calorimetry that measures oxygen uptake and carbon dioxide production to obtain energy expenditure. Oxygen uptake can be translated to kcal (kilocalories) using a constant of 5 kilocalories per liter (kcal$\cdot$L^{-1}), and to kJ (kilojoules) using a constant of 4.19 kilojoules per kilocalories (kJ$\cdot$kcal^{-1}) (U.S. Department of Health and Human Services 1998). The resting metabolic rate is the minimal rate of energy that is necessary for maintenance of basic bodily functions. It is measured following an overnight fast and 8 h of sleep (Howley 2001). To quantify energy expenditure during exercise, the metabolic rate at rest is defined as 1 MET, which has an oxygen uptake of approximately 3.5 milliliter per kilograms per minute (mL$\cdot$kg$^{-1}\cdot$min^{-1}) (U.S. Department of Health and Human Services 1998). Thus, an activity of 7 METs represents an activity that requires seven times the resting metabolic rate. The net energy cost of an activity can be obtained by subtracting 1 MET (RMR) from the gross energy cost (TEE) (Howley 2001).

49.3.2 Dose of Physical Activity: Frequency, Duration, and Intensity

The dose or volume of the energy expended during physical activity represents the product of three components, including the frequency and duration of the activity and the intensity with which the activity is performed (Fig. 49.1). For certain health-related questions, the volume of all-encompassing energy expended may be appropriate. However, in terms of dose-response relationships, the term "physical activity" should be used according to its specific components. For example, in the 1990s interest emerged in whether moderate-intensity activity is sufficient for the promotion of health and in whether vigorous-intensity activity provides further benefit (Pate et al. 1995). This can be considered analogously to the term "diet" in nutritional epidemiology: diet is an overarching term comprising many nutrients and foods. In terms of research questions related to health outcomes, such as cancer or heart disease, one may not only look at diet in its entirety but also at individual dietary factors, such as *trans*-fatty acids or vegetables and fruits consumed (Lee et al. 2009).

In particular, *frequency (how often)* measures the number of physical activity events during a specific time period, for example, per day, per week, or per month (Livingstone et al. 2003). *Duration (for how long)* describes the amount of time spent participating in a physical activity event and is typically expressed in hours or minutes. *Intensity (how hard a person works to perform the activity)* refers to the level of effort required to perform a specific activity. It is expressed in various ways depending on the domain (e.g., recreational activity or occupational activity) and on the time frame within which the activity is performed (i.e., 20 to 60 min for a workout vs. 8 h for a workday).

The *absolute intensity* of physical activity describes the actual rate of energy expenditure. It is commonly expressed in kcal or kJ per minute, oxygen uptake ($\mathrm{L{\cdot}min^{-1}}$), oxygen uptake relative to body mass ($\mathrm{mL{\cdot}kg^{-1}{\cdot}min^{-1}}$), or METs. Alternatively, to the MET definition of 3.5 $\mathrm{mL{\cdot}kg^{-1}{\cdot}min^{-1}}$, 1 MET is also equivalent to an energy expenditure of 1 $\mathrm{kcal{\cdot}kg^{-1}}$ body weight per hour (U.S. Department of Health and Human Services 1998; Howley 2001). Physical activity is often expressed as MET-time/week (MET-hours/week or MET-minutes/week). That variable is created by multiplying the energy cost in MET for each activity by the amount of time (in minutes or hours) spent performing that specific activity during a given week (Lee et al. 2009). Ainsworth et al. (2011) recently provided an updated compendium of MET values that refers to different types of physical activities (Ainsworth et al. 2011). To facilitate research regarding dose-response relationships, activities are grouped into light-intensity activities (1.6 to 2.9 METs) (Patel et al. 2008; Tudor-Locke et al. 2009), moderate-intensity activities (3.0 to 5.9 METs), and vigorous-intensity activities (≥6 METs) (Table 49.1) (Ainsworth et al. 2011).

Relative intensity refers to "the relative percentage of maximal aerobic power that is maintained during exercise" (Fletcher et al. 2001). Persons vary in their fitness status and may respond differently to an exercise performance with a predetermined absolute intensity. To account for such variation, relative intensity has been related to some maximal degree of physiological response. Relative intensity can be

Table 49.1 Classification of physical activity intensity according to Ainsworth et al. (2011)

Classification	Metabolic equivalent tasks (METs)	Examples
Sedentary behavior	1.0–1.5	TV viewing, playing video games
Light intensity	1.6–2.9	Fishing, golf, washing dishes, food shopping
Moderate intensity	3.0–5.9	Easy swimming, yard work, bowling, surfing, yoga
Vigorous intensity	≥6.0	Jogging, tennis, skiing, running, aerobics

expressed as percentage of $VO_{2\,max}$, percentage of oxygen uptake reserve ($\%VO_{2R}$), percentage of heart rate reserve (%HRR), percentage of maximum heart rate ($\%HR_{max}$), or Borg Rating of Perceived Exertion (RPE) (Howley 2001). Borg (1998) developed the RPE which is based on a scale from 6 to 20 to describe a person's perception of exertion during exercise (Borg 1998). In adults, an RPE of 12 to 13 represents moderate intensity of physical activity, which corresponds to 40% to 60% of $VO_{2\,max}$ (Fletcher et al. 2001).

49.4 Descriptive Epidemiology of Physical Activity

The descriptive epidemiology of physical activity helps identify public health priorities and research needs, and it often serves as a basis for the development of national policies and provides guidance to physical activity programs (Lee et al. 2009; Hallal et al. 2012). Information about physical activity behavior in populations is typically gathered using physical activity surveillance systems, which have become a necessary tool in public health. Data on the distribution of physical activity in representative population samples are generally derived from national governmental surveillance systems. The World Health Organization (WHO) regularly provides data on physical activity of representative samples at the global population level (World Health Organization 2012a). While public health surveillance systems that track physical activity prevalence have increased nationwide, only few are well established and monitor physical activity trends on a regular, long-term basis (Hallal et al. 2012).

The current section provides information on the prevalence of physical activity among adults, adolescents, and children. It also describes time trends of habitual physical activity in those populations.

49.4.1 Prevalence and Trends in Physical Activity Among Adults

The Behavioral Risk Factor Surveillance System (BRFSS), the National Health and Nutrition Examination Survey (NHANES), and the National Health Interview Survey (NHIS) are well-established surveillance systems that provide data on

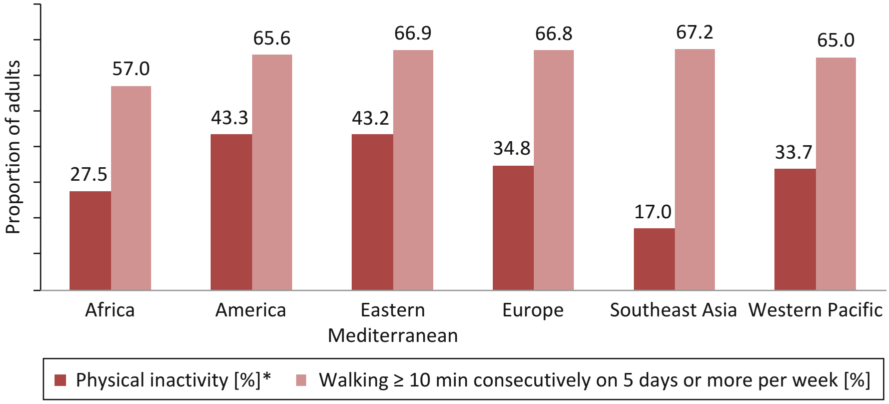


Fig. 49.4 Physical inactivity and walking behavior in individuals aged ≥15 years by World Health Organization region (Hallal et al. 2012); *Physical inactivity was defined as not meeting any of the following criteria: 30 min of moderate-intensity physical activity on at least 5 days every week, 20 min of vigorous intensity physical activity on at least 3 days every week, or an equivalent combination achieving 600 MET-minutes per week

physical activity behavior of the U.S. population (Centers for Disease Control and Prevention 2012). The most recent data from the BRFSS from 2009 reveal that 50.6% of U.S. adults engage in at least 20 min of vigorous physical activity on 3 or more days per week or in at least 30 min of moderate physical activity on 5 days or more per week. Moreover, 75.8% of U.S. adults affirmed the question "During the past month, did you participate in any physical activities?" (Centers for Disease Control and Prevention 2009).

Recently, the Lancet Physical Activity Series Working Group published current data on the prevalence and trends in global physical activity using data from the WHO global health observatory data repository (Hallal et al. 2012). The analysis included self-reported estimates of physical activity from persons aged 15 years or more across 122 countries. Physical inactivity was assessed using the International Physical Activity Questionnaire (IPAQ 2012) and the Global Physical Activity Questionnaire (GPAQ) and was defined as not reaching any of the following criteria: 30 min of moderate-intensity physical activity on at least 5 days per week, 20 min of physical activity of vigorous intensity on at least 3 days per week, or a combination of moderate- and vigorous-intensity physical activity up to 600 MET-min per week. The results show that 31.1% of adults worldwide are physically inactive. The frequency of physical inactivity varies considerably among the different WHO regions: 43.3% of persons are inactive in America, 43.2% in the Eastern Mediterranean, 34.8% in Europe, 33.7% in the Western Pacific, 27.5% in Africa, and 17.0% in Southeast Asia (Fig. 49.4). Estimates across countries vary by sex but in general, women are more inactive (33.9%) than men (27.9%). Also, physical inactivity increases with age (Hallal et al. 2012). On a worldwide basis, 64.1% of adults participate in walking for at least 10 min on 5 or more days per week, with only modest variation between WHO regions (Fig. 49.4). The prevalence of walking differs only weakly between men and women or between age groups. Worldwide, 31.4% of adults report engaging in vigorous-intensity physical activity on 3 or more days per week, with substantial differences between WHO regions. For example, only 25.4% of adults in Europe and 24.6% in America perform vigorous physical activity on at least 3 days per week, compared to 43.2% of adults in the Eastern Mediterranean and Southeast Asian regions. Men are more likely to engage in vigorous-intensity physical activity than women, and physical activity participation decreases with age (Hallal et al. 2012).

Temporal trends of physical activity data show a decline in overall physical activity participation in the past decades and that trend is stronger in industrialized than non-industrialized countries (Brownson et al. 2005; Hallal et al. 2012). Occupational and transportation physical activity have been continuously decreasing, while recreational physical activity has been concurrently increasing over time (Brownson et al. 2005; Knuth and Hallal 2009; Church et al. 2011). For example, Borodulin et al. (2008) examined 30-year time trends in recreational, occupational, and transportation physical activity from 1972 to 2002 across birth cohorts in Finnish adults. They used data from the FINRISK Study, which monitors levels of chronic disease risk factors in Finland. Between 1972 and 2002, the prevalence of recreational physical activity increased from 49% to 67% in women and from 66% to 77% in men, whereas the prevalence of occupational physical activity declined from 47% to 25% in women and from 60% to 38% in men. Moreover, transportation physical activity decreased from 34% to 22% in women and from 30% to 10% in men (Borodulin et al. 2008). Increasing industrialization and emerging technologies along with less physically demanding jobs have led to people becoming more sedentary in their daily routines. Although data are sparse, it has been documented that in parallel with declines in adult physical activity levels, sedentary time has increased, which is attributable to increased television watching and computer use (Brownson et al. 2005; Proper et al. 2011).

Recently, Bauman et al. (2011) reported on sedentary time among adults aged 18 to 65 years from 20 countries. The median sitting time was 300 min (interquartile range, 180 to 480 min) per day. Adults aged 18 to 39 years were more likely to fall into the highest quintile of sitting than adults aged 40 to 65 years (Bauman et al. 2011). Objectively measured physical activity and sedentary behavior data obtained from participants of the Australian Diabetes, Obesity, and Lifestyle (AusDiab) study revealed that on average, adults spent more than half of their daily time in sedentary behaviors, with only 5.8 h per day spent in light-intensity activity and 0.6 h per day spent in moderate to vigorous activity (Healy et al. 2007) (Fig. 49.5). The trend toward an increase in sedentary time and a decline in physical activity might continue to unfold throughout future decades. Ng and Popkin (2012) calculated predictions for physical activity levels in five countries for the years 2020 and 2030, assuming that trends will remain linear. According to that report, in the U.S. in 2009, total physical activity levels were 160 MET h per week and are forecasted to decrease to approximately 142 MET h per week by 2020. Great Britain, China, and Brazil are predicted to achieve that level by the year 2030, whereas in India, total physical activity levels are expected to decline less rapidly (Ng and Popkin 2012).

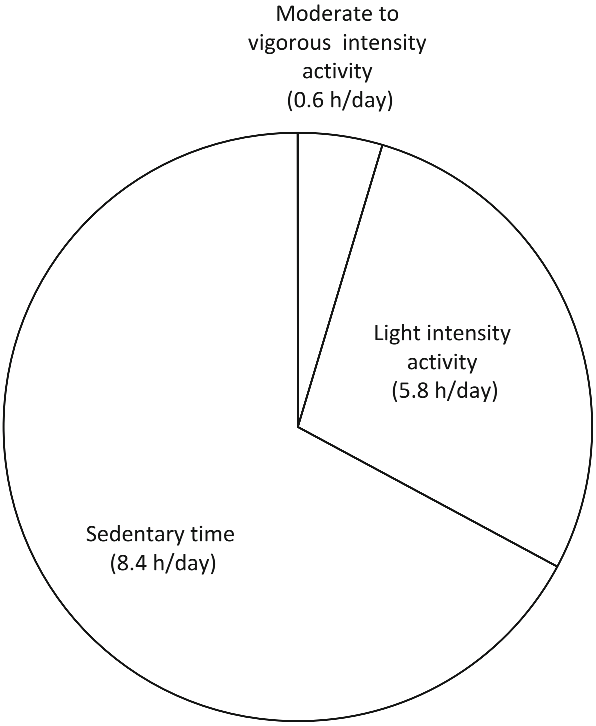


Fig. 49.5 Distributions of sedentary time, physical activity of light intensity, and physical activity of moderate to vigorous intensity in Australian adults (Healy et al. 2007)

49.4.2 Prevalence and Trends in Physical Activity Among Children and Young Persons

Physical activity of children and young persons encompasses sports, games, play, recreation, transportation, physical education, or planned exercise, in family, school, and community settings (World Health Organization 2011). Compared to adults, physical activity patterns in children are sporadic, and continued periods of moderate or vigorous physical activity are infrequent (Armstrong and Welsman 2006). Representative data on physical activity in youth can be derived from the Global School-based Student Health Survey (GSHS) and the Health Behavior in School-aged Children (HBSC) Survey. The Youth Risk Behavior Surveillance System (YRBSS) provides comprehensive data on physical activity prevalence in U.S. youth. The survey is conducted biennially and represents data from 9th to 12th grade students (usually aged 15 to 18 years) in schools throughout the U.S. (Centers for Disease Control and Prevention 2011). The most recent national overview of physical activity from the YRBSS concludes that 13.8% of students do not engage in “at least 60 minutes of any kind of physical activity that increased their heart rate and made them breathe hard some of the time on at least one day during the seven days before the survey." On average, 28.7% of students reported being physically active for at least 60 min on all 7 days (Centers for Disease Control and Prevention 2011). Current data from the GSHS and HBSC included 105 countries worldwide and revealed that 80.3% of 13- to 15-year-olds fail to achieve 60 min of physical activity per day (Hallal et al. 2012).

Data are consistent in that physical activity levels of both sexes are high during childhood and decline thereafter through teenage time. Also, differences exist in physical activity habits among girls and boys. Nader et al. (2008) examined patterns of moderate to vigorous activity among young persons who were followed from ages 9 to 15 years and observed that physical activity levels decreased significantly during that age period (Nader et al. 2008). At the age of 9 years, children engaged in 3 h of moderate- to vigorous-intensity physical activity per day on both weekends and weekdays. By the age of 15, adolescents participated in moderate to vigorous activity for only 49 min per weekday and 35 min per weekend day (Nader et al. 2008).

In a review of habitual physical activity data of European children and adolescents, Adams (2006) stressed the difference in physical activity patterns between sexes, with boys being more engaged in vigorous activity than girls (Adams 2006). Also, the Lancet Physical Activity Series Working Group found girls to be less physically active than boys. The group reported that 80% or more of boys in approximately half of the 105 countries investigated and 80% of girls in 95% of the countries did not reach the recommended 60 min per day of physical activity (Hallal et al. 2012).

As mentioned above, time trends toward lower physical activity levels have been observed in the adult population, and this is also true for children and adolescents. In an analysis of trends in physical activity levels in adolescents according to sex, age, and race, the weighted percentage of adolescents defined as being active was higher in school grades 9 and 10 at all years compared to school grades 11 and 12. A decline in physical activity was observed among all adolescents of school grades 9 and 10 in 2003 compared to 1993. When change in physical activity levels according to sex and year was considered, a downward trend was observed only in boys and only in school grades 9 and 10 (Adams 2006). Recently, a study characterized temporal trends in free time and organized physical activity participation of U.S. youth across different age groups during a 5-year period (Wall et al. 2011). The prevalence of free-time physical activity participation declined linearly from ages 9 to 17 years in both sexes. Organized physical activity declined after the age of 14 years in both sexes. Free-time physical activity participation was lower in girls compared to boys between the ages of 12 and 16 years (Wall et al. 2011). Guthold et al. (2010) used data from the GSHS to analyze sedentary behavior in youth from 32 countries. They demonstrated that in more than half of the countries examined, one third or more of students spend 3 h or more per day in sedentary behaviors. Figures 49.6 and 49.7 illustrate comparisons of weekday time spent in sedentary behavior and moderate- to vigorous-intensity physical activity in 9- and 15-year-olds from four European countries (Nilsson et al. 2009). Across all four countries,

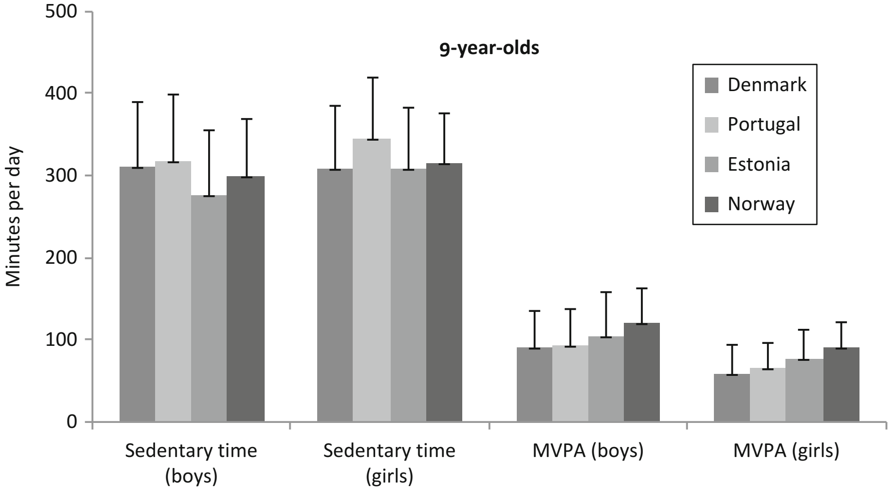


Fig. 49.6 Sedentary time and physical activity levels of moderate to vigorous intensity (MVPA) on an average weekday among 9-year-olds according to sex and country (Nilsson et al. 2009)

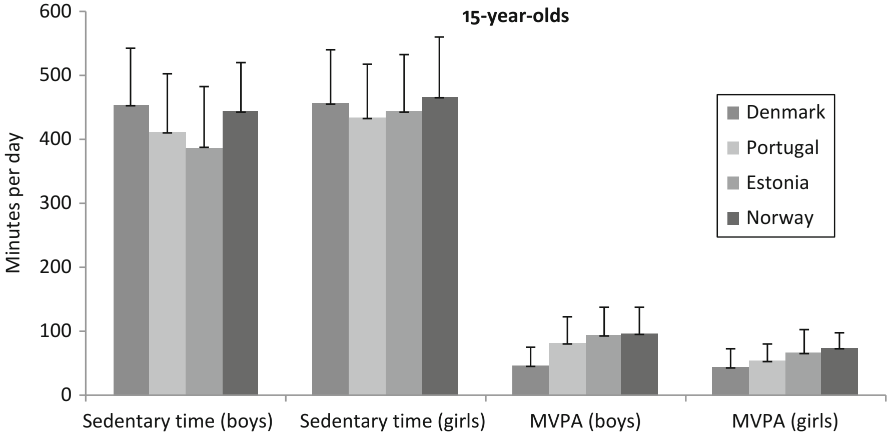


Fig. 49.7 Sedentary time and physical activity levels of moderate to vigorous intensity (MVPA) on an average weekday among 15-year-olds according to sex and country (Nilsson et al. 2009)

15-year-old adolescents spent more time in sedentary behavior than 9-year-old children. In both age groups, boys were more likely to engage in moderate to vigorous activity than girls (Nilsson et al. 2009). Current data from the YRSS revealed that 31.3% of U.S. students played video or computer games or used computers outside school time for 3 or more hours per day on an average school day. A similar amount of students (32.4%) reported watching television 3 or more hours per day on an average school day (Centers for Disease Control and Prevention 2011). Those data underpin the rise in sedentary activities along with an increased time spent in television watching and computer use, a trend that is more pronounced in teens than in children.

49.5 Domains of Physical Research

Physical activity epidemiology includes a number of interrelated research areas, each of which has specific issues and challenges. While research within these specific domains is frequently conducted independently by individual research groups, the body of evidence generated in each research group impacts upon the other research areas. The following section provides a summary of the different domains of research in physical activity epidemiology.

49.5.1 Surveillance Research

Physical activity surveillance research focuses on detecting trends and patterns in physical activity in the population (Hallal et al. 2012). Examples of surveillance research include the development of valid and reliable methods to monitor physical activity patterns in national surveys and the improvement of techniques for documenting long-term physical activity patterns in the population as a whole and among population subgroups, such as children, adolescents, or people with disabilities. This research provides the data needed to prioritize public health activities, identify research needs, allocate funding, inform policy to promote the public's health, direct physical activity promotion efforts, and identify appropriate behavioral intervention targets aimed at increasing physical activity levels in the population. Physical activity measures for this area of research depend almost entirely on self-reports or interviewer-administered assessments because they can be applied to large groups of the population at low cost. A critical issue in physical activity surveillance research is the need for high accuracy and precision of the assessment instruments because small inconsistencies in the wording of questions can translate into considerable variation in estimates of the prevalence of physical activity. Such heterogeneity would make it challenging to track physical activity patterns over time or to compare prevalence estimates of physical activity across subgroups of the population. Surveillance research also encompasses the area of correlates research, which studies the determinants of physical activity through the identification and multilevel modeling of personal, social, economic, and environmental determinants of physical activity (Trost et al. 2002). Physical activity correlates research identifies mediators that can be used to inform physical activity intervention research.

49.5.2 Health Outcomes Research

Health outcomes research is aimed at exploring the associations between physical activity and health (Adami et al. 2010). Examples of this type of research include investigating the association of physical activity with heart disease, diabetes, cancer, bone loss, or quality of life or studying the health benefits of physical activity in specific subgroups of the population, such as racial and ethnic minorities, the elderly, or people with disabilities. Physical activity assessments for this research area vary depending on the particular research question but are primarily based on self-reported questionnaires due to the large sample sizes typically needed. Self-reported instruments usually provide a simple ranking of physical activity levels based on job titles or time spent in recreational activities. Despite their imperfect nature, self-reported measures of physical activity have been successfully employed in numerous large epidemiological studies and have provided evidence of important health benefits of physical activity. Lack of precision may lead to underestimation of the health effects of physical activity. Although simple classifications have been adequate to uncover strong associations between physical activity and disease risk in the past, more refined physical activity assessments will be needed in the future to uncover more complex relations of physical activity with risk of disease, such as the delineation of the relation of specific dimensions of physical activity (e.g., duration, frequency, intensity) to disease risk.

49.5.3 Physical Activity Intervention Research

Physical activity intervention research is aimed at creating and evaluating programs designed to promote physical activity. Focus is often placed on population subgroups that have been found in surveillance research to be at increased risk for sedentary behavior, such as lower socioeconomic populations or children and adolescents (Heath et al. 2012). The principal goal in designing an appropriate intervention program is to organize and prioritize the numerous factors that are known to impact upon physical activity and to integrate them in an intervention program. The effectiveness of physical activity interventions is greater when established behavioral theory is taken into consideration in designing the intervention. The targets of physical activity interventions have evolved considerably over the past several decades as physical activity recommendations and principles of behavioral modification have developed. Because early physical activity guidelines emphasized vigorous forms of physical activity for health promotion, initial intervention programs were directed at structured exercise and cardiorespiratory fitness. By comparison, current physical activity recommendations focus on moderate-intensity activity and on avoidance of sedentary behavior as means of promoting health. Thus, current interventions are aimed at facilitating the development of individualized lifestyle physical activity programs that enable people to accumulate the recommended dose of moderate-intensity physical activity over the day. Because an individual's behavior is strongly determined by the social, economic, and physical environment, physical activity interventions often target multiple domains and levels of influence with the rationale of enhancing the probability of behavior modification (Kahn et al. 2002).

49.5.4 Mechanistic Physical Activity Research

Mechanistic or basic research on physical activity involves any research discipline that aims at elucidating the etiological pathways by which physical activity influences the human organism, including the biological mechanisms through which physical activity impacts upon health. Results from basic research are often used as the basis for formulating research hypotheses in health outcomes research and physical activity intervention research. One illustration of the interrelationship between mechanistic physical activity research and health outcomes research is provided by the link between the health benefits related to regular physical activity. Numerous epidemiological studies have found that an increased level of physical activity decreases the risks of certain types of cancer, such as colon cancer (Wolin et al. 2009). Basic science research investigating potential biological mechanisms suggests that high levels of physical activity may reduce colon cancer risk through reductions of adipose tissue, insulin resistance, chronic low-grade inflammation, or oxidative stress (Quadrilatero and Hoffman-Goetz 2003). The advancement of knowledge in the area of health outcomes research is closely related to continuous progress in basic science research.

49.6 Physical Activity Recommendations

Starting in the 1960s, recommendations for physical activity participation based on scientific data were disseminated by health and fitness organizations for the purpose of promoting health. In addition, several organizations and government agencies drew upon physical activity recommendations designed to address specific health issues such as the attainment of weight control, cardiorespiratory and muscular fitness, and cancer prevention (Garber et al. 2011). A detailed review of various physical activity guidelines is beyond the scope of this chapter. However, we will provide a brief summary of the main landmarks in the development of recommendations for the promotion of physical activity.

In 1965 and 1972, the President's Council on Physical Fitness and the American Heart Association (AHA) released the first guidelines for physical activity programs and exercise prescription as means of improving physical activity performance and promoting health (U.S. Department of Health and Human Services 1998). In 1975, the American College of Sports Medicine (ACSM) issued the guidelines entitled "Graded Exercise Testing and Prescription" followed by the release of "The Quantity and Quality of Exercise for Developing and Maintaining Fitness in Healthy Adults" in 1978. Those recommendations focused on cardiorespiratory fitness and body composition, endorsing exercise training of 3 to 5 days per week with durations of 15 to 60 min per training session and an exercise intensity of 50% to 85% of VO_{2R} and the use of large muscle groups. In 1990, the ACSM updated its position statement, aiming at muscular strength, endurance, and flexibility of the major muscle groups that can be reached by a training program including resistance training and flexibility exercises (U.S. Department of Health and Human Services 1998).

Once it became evident that activities of moderate intensity provide health benefits independently of cardiorespiratory fitness, in 1995 the ACSM and the Centers for Disease Control and Prevention (CDC) coincidently published physical activity guidelines recommending that "every adult should accumulate 30 or more minutes of moderate intensity on most, preferably all, days of the week" (Pate et al. 1995). In addition to the ACSM/CDC (1995), the Institute of Medicine (IOM) of the National Academy of Sciences, the Office of the U.S. Surgeon General (OSG), and the U.S. Department of Health and Human Services (HHS)/Department of Agriculture concertedly advocated that engaging in moderate activity for at least 30 min on most days of the week might produce health benefits (OSG 1996; ACSM/CDC 1995; Pate et al. 1995; U.S. Department of Health and Human Services 1996; U.S. Department of Health and Human Services and U.S. Department of Agriculture 2000; Institute of Medicine of the National Academy 2002).

In order to respond to the population's increasing engagement in cardiorespiratory fitness and resistance training programs and frequent queries for exercise prescription, in 1998 the ACSM provided guidelines for the development and maintenance of cardiorespiratory fitness, body composition, muscular strength and endurance, and flexibility in healthy adults (Pollock et al. 1998).

In 2002, the IOM Committee on Dietary Reference Intake published a report that included the recommendation of 60 min of daily physical activity to prevent unhealthful weight gain (Institute of Medicine of the National Academy 2002). In 2003, the International Association for the Study of Obesity (IASO) addressed the continuously increasing obesity problem by advocating 40 to 50 min of physical activity per day for the prevention of obesity and 60 to 90 min per day for the prevention of weight regain in formerly obese persons (Saris et al. 2003).

In accordance with the 1995 ACSM/CDC physical activity guidelines, a large body of evidence supported the recommendation of lifestyle activities as health-enhancing physical activity (HEPA) (Oja et al. 2010). The term HEPA refers to the recommendation of a minimal volume of physical activity for the prevention of various diseases and premature death and for the enhancement of health (Pate et al. 1995; Bucksch and Schlicht 2006). It does not only include recreational physical activity but also incorporates occupational and household activities that can represent planned or unplanned activities that fit best into everyday life (Dunn et al. 1998). Bucksch and Schlicht (2006) reviewed the epidemiological evidence on the HEPA recommendation. They specifically addressed the questions of whether the accumulation of 30 min of moderate-intensity physical activity per day is sufficient and whether the health-enhancing effect is independent of the domain of physical activity (transportation, recreation, and household) and dimension (frequency, duration, and intensity) of physical activity. They concluded that moderate physical activity of 30 min per day, independent of the domain, provides increased health benefits in sedentary men and women (Bucksch and Schlicht 2006). The HEPA Europe, a collaborative project initiated in 2005, aims to improve health through physical activity among all persons in the WHO European region by supporting and strengthening efforts to increase physical activity participation and by improving the conditions for healthy lifestyles. This includes the development and dissemination of effective strategies, programs, and other examples of good practice for the promotion of HEPA and seeks to facilitate and support multi-sectoral approaches to the promotion of HEPA (World Health Organization 2005).

In 2007, the ACSM and AHA jointly published an updated physical activity recommendation. This update still recommends 30 min of moderate-intensity activity participation but specifies that physical activity be performed for "at least five days per week" rather than "at most, preferably, all days." Physical activity participation may be performed in an accumulation of 10-min bouts (Haskell et al. 2007). The guideline also included an additional statement that 20 min of vigorous intensity on 3 days per week might substitute the 30 min of moderate-intensity physical activity on at least 5 days per week. The recommendation also states that a combination of vigorous- and moderate-intensity activities may produce health benefits.

Also, in 2007 the ACSM issued a separate recommendation for older people aged 65 years or older and for persons aged 50 to 64 years with clinically significant chronic conditions and/or functional limitations (Nelson et al. 2007), which was subsequently updated in 2009 (Chodzko-Zajko et al. 2009).

The *2008 Activity Guidelines for Americans* published by the HHS encourage physical activity participation for health promotion by stating that "physical activity is better than non-activity and adults who participate in any amount of physical activity gain some health benefits" (U.S. Department of Health and Human Services 2008).

The most recent ACSM recommendations are based on the 1998 ACSM guidelines and were released in 2011 with the purpose of promoting physical fitness and health among adults. The 2011 ACSM guideline recommends that most adults should (Garber et al. 2011):

- Perform moderate-intensity cardiorespiratory exercise training for at least 30 min per day on 5 or more days per week or cardiorespiratory exercise training of vigorous intensity for at least 20 min per day on 3 or more days per week ($\geq$75 min per week) or a combination of moderate- and vigorous-intensity exercise.
- Engage in resistance exercise for each of the major muscle groups on 2 to 3 days per week and perform neuromotor exercise (coordination, balance, and agility).
- Do flexibility exercises for each of the major muscle groups (a total of 60 seconds per exercise) on 2 or more days per week.

The exercise training should be modified according to usual physical activity habits, physical function, health status, and exercise responses of the individual (Garber et al. 2011). Moreover, the recommendations include reducing the total time spent in sedentary behaviors and incorporating frequent, short bouts of standing and physical activity between periods of sedentary engagement.

In 2010, the WHO released their "Global Recommendations on Physical Activity for Health" specifically for adults, older persons, and children for the primary prevention of non-communicable diseases. In particular, those recommendations intend to improve cardiorespiratory and muscular fitness, bone health space and cardiovascular and metabolic health markers and reduce symptoms of anxiety and depression. The WHO guidelines give the following recommendations (World Health Organization 2011):

- Children and young persons aged 6 to 17 years should engage in at least 60 min of moderate- to vigorous-intensity physical activity per day, most of which should be aerobic. At least three times per week vigorous-intensity activities should be performed to strengthen muscle and bone.
- Adults aged 18 to 64 years should spend at least 150 min (300 min for additional health benefits) in moderate-intensity aerobic physical activity over the week, or engage in at least 75 min (150 min for additional health benefits) of vigorous-intensity activity over the week, or should combine moderate and vigorous-intensity activities in an equivalent manner. Aerobic physical activity should be performed in bouts lasting for at least 10 min, and muscle-strengthening activities should involve major muscle groups and should be performed on 2 or more days of the week.
- Adults aged 65 years or older should perform physical activity as described for adults aged 18 to 64 years. Persons with poor mobility should engage in physical activity to enhance balance and to prevent falls on 3 or more days per week.

Although muscular injury is uncommon at the recommended level of 150 min per week of moderate-intensity activity, the WHO recommends starting with a moderate physical activity program and gradually increasing to higher physical activity levels (World Health Organization 2011).

A number of countries such as Australia (Australian Government Department of Health and Ageing 2010), Canada (Canadian Society for Exercise Physiology 2012), and the European Union (EU) Member States (European Commission 2008) have national physical activity recommendations. By mid-2013, the EU Expert Group on Sport, Health and Participation will issue guidelines designed to improve "ways to promote health-enhancing physical activity and participation in grassroots sport, and to identify respective measures" (European Commission 2012).

Several studies have investigated adherence to physical activity recommendations in relation to risk of total mortality. The 1990 ACSM and 1995 ACSM/CDC recommended levels of vigorous and moderate physical activity were investigated in relation to the risk of mortality in a sample of men and women from Germany. A significant inverse association was found between moderate activity at recommended levels and risk of mortality in women but not men. By comparison, an inverse association with mortality was seen at recommended levels of vigorous activity in men, but not women (Bucksch 2005). Leitzmann et al. (2007) examined the effect of engaging in physical activity according to the physical activity guidelines of the ACSM/CDC (1995) and OSG (1996) on all-cause mortality in a large U.S. cohort study (Leitzmann et al. 2007). Following the physical activity guidelines of 30 min of moderate activity on most days of the week or engaging in 20 min of vigorous exercise three or more times per week was associated with 27% and 32% reductions in all-cause mortality, respectively (Leitzmann et al. 2007). Schoenborn and Stommel (2011) studied mortality risk associated with adherence to the 2008 Physical Activity Guidelines for Americans. They concluded that meeting the recommendations was associated with a reduced risk of all-cause mortality among U.S. adults, particularly in those with chronic conditions, with odds ratios (OR) ranging from 0.65 to 0.75 (Schoenborn and Stommel 2011).

49.7 Measurement of Physical Activity

Physical activity is a complex, multidimensional exposure that is difficult to measure accurately among free-living individuals. Several physical activity assessment methods exist, each coupled with specific strengths and limitations. Two main types of methods for measurement of physical activity can be distinguished: subjective methods (e.g., questionnaires, diaries, recalls, and logs) and objective methods (e.g., activity monitors, heart rate monitors, DLW, and indirect calorimetry). Subjective methods provide the poorest validity among physical activity assessment methods. Because of their high accuracy, DLW and indirect calorimetry methods have frequently been considered validation criterion methods for other types of physical activity assessments (Fig. 49.8). Different assessment methods measure distinct dimensions of physical activity. As mentioned above, physical activity represents a behavior, whereas energy expenditure mirrors the energy costs of physical activity behavior. Motion sensors such as accelerometers and pedometers measure physical activity which can be converted to energy expenditure by equations, whereas DLW and indirect calorimetry directly assess energy expenditure. Lamonte and Ainsworth (2001) stated that energy expenditure may be a better predictor of health-related outcomes than the specific type of physical activity that results in the expended energy. Therefore, researchers often convert outputs of physical activity

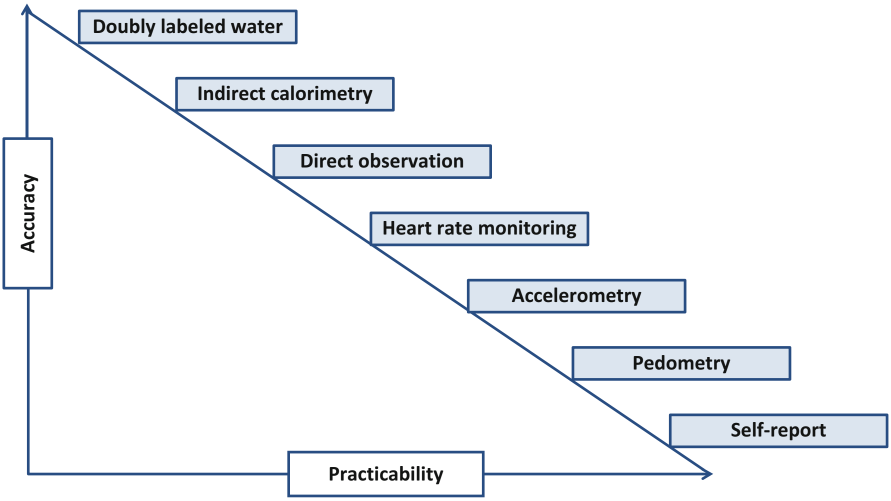


Fig. 49.8 Practicability versus accuracy of different physical activity assessment methods (modified from Mueller et al. (2010))

measurements to units of energy expenditure (Lamonte and Ainsworth 2001). For example, Manson et al. (1999) reported that the association between physical activity and coronary events was similar for both light- to moderate-intensity physical activity and vigorous-intensity exercise when total energy expended was comparable.

While the method for assessment of physical activity should fit the research question, the choice of method will usually represent a trade-off between accuracy and feasibility (U.S. Department of Health and Human Services 1998). The research instruments used must be evaluated not only for their efficacy to assess an individual's physical activity but also for their applicability in large-scale epidemiological settings (U.S. Department of Health and Human Services 1998). Physical activity assessment methods vary in that they are appropriate for specific age groups of the population. For example, direct observation is an assessment method that usually intends to monitor children's physical activity behavior in various settings including school, home, parks, and recreation areas.

Before measuring physical activity in a study, measurement instruments should be tested for reliability and validity in a population which is similar to the one that will be examined. Validity describes how well the instrument assesses the exposure it intends to measure, and reliability assures that the method of measurement delivers the same results if reported under the same circumstances. Both validity and reliability represent quality criteria for measurement instruments.

The following section reviews different methods used for the measurement of physical activity and provides information on their use in different population subgroups (e.g., children, adults, and older persons). Strengths and limitations of the different methods of physical activity assessment will be discussed. An evaluation of instruments for physical activity assessment according to specific attributes is illustrated in Table 49.2.

49.7.1 Subjective Physical Activity Assessment Methods

Subjective methods, also termed self-reported methods, are the most convenient way to collect physical activity data on a large number of individuals. They are most practical for assessing physical activity in large epidemiological studies (Caspersen 1989). Self-reports include questionnaires, recall surveys, diaries, and logs. They have in common that they are inexpensive, relatively easy to administer, and highly acceptable for study participants (LaPorte et al. 1985; Caspersen 1989). However, subjective measurements suffer from poor validity, as outlined below (Vanhees et al. 2005).

49.7.1.1 Self-Reported Physical Activity Questionnaires

Physical activity questionnaires (PAQs) represent self-reported instruments for measuring physical activity that are widely used in epidemiological settings. These instruments can be interview-based or self-administered and usually query physical activity over the past day, week and month but also over a year or even over the

Table 49.2 Evaluation of physical activity assessment instruments according to specific attributes

	Costs	Acceptance by participants	Accuracy	Ease of administration	Measurement of dimensions	Induction of change in PA behavior	Suitable for children (<10 years of age)	Feasible in large-scale free-living populations
Subjective methods								
Self-reported PAQ	++	++	0	++	+	+	–	++
Recall	+	+	0	+	+	+	0	+
Diary	++	0	0	+	++	0	–	0
Log	++	0	0	+	+	0	0	0
Objective methods								
Pedometry	++	++	+	+	–	+	++	+
Accelerometry	0	+	+	+	+	+	+	+
Direct observation	–	+	++	0	++	0	++	0
Heart rate monitoring	+	0	+	0	+	+	–	0
DLW	–	–	++	+	–	+	++	–
Indirect calorimetry	–	–	++	+	–	0	++	–

– poor, 0 acceptable, + good, ++ very good lifetime. PAQs allow the assessment of different components of physical activity, including type, frequency, intensity, and duration in household, occupation, recreation, and transportation. As occupations have become less physically demanding in industrialized countries over the past decades, PAQs have focused on recreational activities above other domains (Lee et al. 2009). To understand the impact of environmental influences on all domains of physical activity, questionnaires have begun to consider physical activity in other domains including transportation, occupation, and household activities (Brownson and Boehmer 2004). While in earlier studies questionnaires queried job titles only (Lee et al. 2009), recent occupational PAQs inquire about physical activity patterns in greater detail including the frequency, duration, and intensity of activity performed by individuals at work (Friberg et al. 2006; Friedenreich et al. 2007; Lee et al. 2009). Questionnaires differ in their type, complexity, and time frame for which physical activity is assessed (Lee et al. 2009). PAQs may be less accurate in assessing activities of light to moderate intensity compared to activities of vigorous intensity (Warren et al. 2010). This may be explained by the fact that low-intensity activities, such as walking, that occur throughout the day are more difficult to assess than structured and planned exercise of vigorous intensity, such as jogging and swimming (Bassett et al. 2000).

The IPAQ was developed in 1996 by Booth and colleagues, and it represents an established population surveillance tool for physical activity assessment to compare physical activities across countries. The IPAQ has been translated into 22 different languages, including Arabic, Croatian, Danish, Dutch (Belgian), English, Estonian, French, German, Icelandic, Korean, Lithuanian, Malaysian, Norwegian, Persian-Farsi, Polish, Spanish (Argentina), Spanish (Columbia), Spanish (U.S.), Swedish, Taiwanese, Turkish, and Vietnamese (International Physical Activity Questionnaire 2012; https://sites.google.com/site/theipaq/questionnaires). It encompasses a set of four questionnaires including a long version (five activity domains asked independently) and a short version (four generic items) for use by either telephone- or self-reported methods (Craig et al. 2003). The GPAQ was developed by the WHO as a physical activity surveillance tool mainly for use in developing countries. It collects information on physical activity engagement at work, travel, and in recreation (World Health Organization 2012b). The European Prospective Investigation into Cancer Study-Norfolk cohort (EPIC-Norfolk), a large population-based cohort study, designed the EPIC Physical Activity Questionnaire 2 (EPAQ2) which measures activity in different domains of life in order to assess energy expenditure (Wareham et al. 2002). The Baecke Questionnaire represents a short assessment tool aimed at estimating physical activity in disease-free populations over the past year (Baecke et al. 1982).

PAQs are principally tailored to assess moderate- to high-intensity physical activity among healthy young to middle-aged adults, but not other age groups, such as the elderly and children. Children's activity is characterized by short bouts of activity rather than more continuous periods of activity (Pate 1993; Sirard and Pate 2001). Also, cognitive immaturity in children or cognitive degeneration in the elderly makes the application of commonly used self-reported questionnaires difficult in those populations. To overcome this obstacle, specific questionnaires for children and elderly people have been designed. They can be self-reported if feasible or proxy-reported by parents or relatives. Examples of PAQs in children are the self-reported and proxy-reported Children's Leisure and Activities Study Survey (CLASS) questionnaires for 10- to 12-year-olds and 5- to 6-year-olds that intend to assess the type, frequency, and intensity of physical activity over a typical week (Telford et al. 2004). Questions such as "Do you usually bounce on a trampoline," "play with pets," or "play indoors with pets?" are asked. The "Motorik-Modul" (MoMo), as part of the German Health Interview and Examination Survey for Children and Adolescents (KiGGS), explores potential associations between motor fitness, physical activity, and health in cross-sectional and longitudinal investigations. Physical activity is assessed by a questionnaire that contains 51 items about the intensity, frequency, and duration of physical activity in everyday life, during recreation time, at school, and in sports clubs. Moreover, a test profile that consists of 11 items is used to determine cardiorespiratory fitness, strength, coordination, and mobility (Opper et al. 2007).

The Yale Physical Activity Survey, the Physical Activity Survey for the Elderly, and the Zutphen Physical Activity Questionnaire have been specifically designed for the elderly (Washburn 2000). In 1991, the Baecke Questionnaire was slightly modified (Modified Baecke Questionnaire) by Voorrips and coworkers to capture habitual physical activity in older populations (Voorrips et al. 1991).

Information attained from PAQs is frequently converted into estimates of energy expenditure (i.e., kcal, kJ, or METs) or some other summary measure that can be used to categorize or rank persons by their physical activity level. Activity codes and MET intensities can be obtained from a physical activity compendium created by Ainsworth and colleagues (Ainsworth et al. 2011). A compendium designed to quantify children's and adolescent's activity levels is also available (Ridley et al. 2008).

However, most PAQs do not accurately estimate energy expenditure (Macfarlane et al. 2006; Neilson et al. 2008). A vast number of criterion-related validity studies and tests of the reliability of PAQs have been conducted. Neilson et al. (2008) evaluated the validity of PAQs for estimation of AAE, using DLW as a criterion method. Only 5 out of 19 PAQs revealed correlation coefficients >0.60 (Neilson et al. 2008). The long and short versions of the IPAQ were validated against accelerometer data in 14 centers across 12 different countries (Craig et al. 2003). The IPAQ questionnaires produced data with a reasonable reliability (Spearman's correlation coefficient $r = 0.8$). The criterion validity provided a Spearman's correlation coefficient of 0.30 (median of the pooled data). These values were comparable to the reliability and criterion validity correlations achieved from the evaluation of seven other self-reported PAQs (Sallis and Saelens 2000).

Boon et al. (2010) assessed the validity of the New Zealand physical activity questionnaire long form (NZPAQ-LF) and the IPAQ-long form (IPAQ-LF) against the accelerometry method. Both the NZPAQ-LF and the IPAQ-LF were only modestly correlated with accelerometer data with respect to moderate physical activity (Spearman's correlation coefficient $r = 0.19–0.30$) and total physical activity (Spearman's correlation coefficient $r = 0.30–0.32$) (Boon et al. 2010). Moreover, the validity of PAQs may vary with obesity status. Norman et al. (2001) evaluated a self-administered PAQ against 7-day activity records in a group of middle-aged and elderly men. They found significantly lower Spearman's correlation coefficients in obese men compared to normal-weight men (Spearman's correlation coefficients $r = 0.39$ vs. $r = 0.73$) (Norman et al. 2001).

An advantage of PAQs is that physical activity levels can be assessed across different domains. Moreover, they are inexpensive and of low burden for study participants and investigators compared to other self-reported instruments such as physical activity diaries or logs. PAQs can easily be administered under free-living conditions to most individuals in the population at a large-scale level. However, activity information obtained by self-reports is subject to recall and social desirability that may influence the validity and/or precision in physical activity measurement. Specifically, validity appears to be poor for low-intensity physical activities (Warren et al. 2010). There is ongoing research regarding the development of more advanced techniques in self-reported approaches, such as Internet-based assessment methods (Schatzkin et al. 2009; cf. Chap. ▶Internet-Based Epidemiology of the handbook). Electronic and internet-based questionnaires may increase standardization in assessment procedures, they are cost and time saving, they are resistant to coding errors, and they allow for an immediate calculation and storage of the data (Warren et al. 2010). Incorporating video guides or spoken instructions rather than paper-based instructions may be of advantage to individuals with poor literacy (Ridley et al. 2001). Researchers using a questionnaire must be aware of the primary outcome of the questionnaire (e.g., the dimensions and domains of activity the questionnaire is designed to measure) and ensure that the questionnaire fits with the research question. Box 49.1 provides a list of questions they may be considered before conducting a study.

Box 49.1. Possible considerations before using a questionnaire in a physical activity study (Adapted from Warren et al. (2010))

1. Does the questionnaire fit the research question?
2. Does the questionnaire adequately address the population under investigation?
3. What is the time frame for which physical activity is assessed (e.g., past week, month, year)?
4. What types of physical activity does the questionnaire capture?
5. Which dimensions of physical activity does the questionnaire include?
6. How will the questionnaire be administered (self-administered or interviewer-administered)?
7. Does the questionnaire enable to clearly follow the instructions for unmistakably responding to the questions, or in interview-based assessments, are the interviewers well trained?
8. Have tests for reliability been undertaken?

9. Has the questionnaire been validated against objective physical assessment instruments?
10. How will the data be stored and analyzed, and how will outliers and missing values be treated?

49.7.1.2 Retrospective Quantitative History

The retrospective quantitative history represents the most complex form of recall surveys and provides details on past physical activity details typically over 1 year (Caspersen 1989; U.S. Department of Health and Human Services 1996) or even a lifetime (Lee et al. 2009). These surveys are interviewer-based or self-administered. Recall of historical activity is useful to determine the relationship between physical activity and chronic disease that have a long developmental period, such as osteoporosis or cancer (Winters-Hart et al. 2004; Lee et al. 2009) and to establish physical activity and health outcome relationships in studies where long-term or lifetime patterns of physical activity play an important role (Bowles et al. 2004).

Winters-Hart et al. (2004) tested the validity of physical activity data derived from a historical physical activity questionnaire (HPAQ) by comparison with actual PAQ data collected at four points in time over a 17-year period in postmenopausal women. Specifically, in the year 2000, participants recalled their physical activity participation from the years 1982, 1985, 1995, and 1999. At each of the points in time studied, significant correlations in the range of 0.39 to 0.62 between the HPAQ data and the data from the actual PAQ were found, with the highest Spearman's correlation coefficients reached at the most recent points in time (Winters-Hart et al. 2004). In the Medical Research Council Ely Study, Besson et al. (2010) validated an HPAQ against objectively measured physical activity from the same time periods up to 15 years ago (Besson et al. 2010). In their study, physical activity data were obtained using calibrated heart rate monitoring during two periods (between 1994 and 1996 and between 2000 and 2002) and an HPAQ administered to assess physical activity from the age of 20 years to the current age. For both time periods, modest correlations between HPAQ-derived and objectively assessed total AEE and vigorous physical activity were found (Spearman's correlation coefficients $r = 0.44$ and 0.40, respectively) (Besson et al. 2010).

Although the HPAQ provides a comprehensive amount of data, it has the disadvantage of representing a considerable burden to the participant to recall physical activity over a long period of time. Moreover, this method is related to a sizeable administrative burden (Lee et al. 2009; U.S. Department of Health and Human Services 1998).

49.7.1.3 Physical Activity Recall Instruments

Physical activity recall instruments attempt to query a wide range of activities over a defined time frame ranging from 1 week to a lifetime. Either they ascertain precise details about physical activity including type, frequency, and duration or they assess usual or typical participation in physical activities (LaPorte et al. 1985;

Blair et al. 1991). The participant is typically asked to recall past activities, during an interview, in person or by phone. The Stanford 7-Day Recall (7-DR) and the Adolescent Physical Activity Recall Questionnaire (APARQ) provide examples of recall questionnaires that collect information on physical activity over the past 7 days (Sallis et al. 1985; Booth et al. 2002). In short-term physical activity recalls, unannounced telephone-administered interviews are performed, and the participant is asked to recall the amount of time spent during the past 24 h or longer (up to 1 month) in physical activities in occupation, household, and recreation (Hu 2008). Because of daily and weekly variation in physical activity, a number of recalls (at least two) may be necessary to assess long-term physical activity patterns.

Richardson et al. (2001) assessed the reliability and validity of the Stanford 7-DR within the Survey of Activity, Fitness, and Exercise (SAFE) study by comparing the 7-DR with physical activity records, accelerometer readings, $VO_{2\,max}$, and percent body fat. Study participants performed 14 recalls over a 12-month period, one recall approximately every 26 days. The validity of the 7-DR was measured at study visits 10 and 11. In terms of reliability, the authors reported age-adjusted sex-specific Spearman partial correlation coefficients of $r = 0.60$ and $r = 0.36$ for men and women, respectively. 7-DR and physical activity record measures were significantly correlated in men (Spearman partial correlation coefficients $r = 0.58$ at visit 10 and $r = 0.66$ at visit 11), whereas in women, a weaker relationship was observed (Spearman partial correlation coefficients $r = 0.32$ at visit 10 and $r = 0.33$ at visit 11). Comparisons of the 7-DR with $VO_{2\,max}$ and percent body fat showed less consistent correlations with $VO_{2\,max}$, but more consistent relationships for percent body fat in women than men. Data from the 7-DR were significantly related to accelerometer data in men (Spearman partial correlation coefficient $r = 0.54$ at visit 10 and $r = 0.45$ at visit 11) but not women (Spearman partial correlation coefficient $r = 0.20$ at visit 10 and $r = 0.06$ at visit 11) (Richardson et al. 2001).

Because of its retrospective nature, recall methods generally do not influence the individual's activity behavior, and they tend to be less burdensome to the participant and the investigator compared to diaries or logs. However, they require memory of past activities which may lead to recall errors (Baranowski 1985; U.S. Department of Health and Human Services 1998). Additional limitations of the method include the time and costs for training the interviewers. Also, recall questionnaires require an individual's cooperation to remember physical activity participation. Moreover, recall questionnaires are less suitable to assess physical activity in children and the elderly.

49.7.1.4 Physical Activity Diaries

Physical activity diaries (also known as physical activity records) usually represent paper and pen-based instruments. The study participants record all their activities as they typically occur during a 24-h period which is usually broken down into 15-min time intervals. The participant can choose specific activities from a predefined list, but space is provided for other activities. Reported activities are then processed using coding schemes which classify each activity by its MET value and thus allow estimating intensity of physical activity. Physical activity diaries are commonly limited to time spans of 1 to 3 days (Caspersen 1989).

When compared to the indirect calorimetry method, physical activity diaries have demonstrated acceptable accuracy in the measurement of daily energy expenditure (Warren et al. 2010).

Advantages of physical activity diaries are that they provide detailed and comprehensive information about the physical activities undertaken, including bouts and patterns of physical activity performed during a day across all domains. The method is cheap to administer and it does not rely on recall or memory (LaPorte et al. 1985; Caspersen 1989). A disadvantage is that it is cumbersome for participants to keep the diary as it requires a detailed recording of the individual's physical activity which is a time-consuming and tiresome task demanding good cooperation and compliance of the participant. Also, the need to record physical activity may lead to a change in an individual's activity behavior, also known as a reactivity effect (Matthews 2002). Physical activity diaries consume a considerable amount of time, and they impose complex data processing on the investigator (e.g., decoding the activity entries) (LaPorte et al. 1985; Haskell and Kiernan 2000). Physical activity diaries represent the most suitable subjective method to estimate energy expenditure in epidemiological studies. However, the recording and analysis process is time-consuming and inconvenient for both the study participant and the investigator. Thus, this method is often used as a reference method for the validation of self-reported PAQs. For example, the validity of the PAQ used in the Nurses' Health Study was assessed by comparison with 4 past-week physical activity recalls and four 7-day physical activity diaries, administered every 3 months (Wolf et al. 1994).

49.7.1.5 Physical Activity Logs

In contrast to physical activity diaries, where bouts of activity are recorded, in physical activity logs, the participant records broad categories of activities (inactive, sitting; light, moderate, vigorous, and very vigorous activity), usually in a structured form (Bouchard et al. 1983). Similar to physical activity diaries, logs provide a considerable amount of physical activity information, albeit not as detailed as diaries. Macfarlane et al. (2006) compared six physical activity measurement methods (triaxial/uniaxial accelerometers, pedometer, heart rate monitoring, physical activity log, and PAQ) and found weak to modest correlations between the physical activity log and other physical activity measurement methods. Specifically, correlations with a triaxial accelerometer (Tritrac) yielded correlation coefficients of $r = 0.23$ for light activity, $r = 0.32$ for moderate activity, and $r = 0.22$ for vigorous activity (Macfarlane et al. 2006). Schmidt et al. (2003) reported that logs were kept more accurately in leaner women compared to those who were overweight.

Advantages of physical activity logs are that they provide detailed data on physical activity patterns of individuals. Furthermore, physical activity logs are less time-consuming for the researcher to process than activity diaries, making them particularly useful for validation studies. They are not subject to bias due to the cognitive demands of the recall process. However, limitations of physical activity logs are the potential for loss of information for activities not represented on the list (e.g., spontaneous activity). Also, logs require some degree of literacy of the participant, and the recording process may produce changes in activity behavior. As with diaries, the recording and data process of logs can be inconvenient for the participant and investigators, and the recording process may produce changes in activity behavior (U.S. Department of Health and Human Services 1998).

49.7.2 Objective Physical Activity Assessment Methods

An alternative to self-reported physical activity assessment methods is the direct measurement of physical activity through mechanical tools, electronic instruments, direct observation, or physiological measurements. These approaches eliminate the obstacles of limitations in recall and self-reporting, and instruments directly provide an accurate assessment of energy expenditure. However, some of these methods are themselves limited by a high burden on the participant and by high costs, or they are restricted to use in laboratory settings only. Therefore, these measurement tools have predominantly been used in small-scale studies (U.S. Department of Health and Human Services 1998). DLW and indirect calorimetry are frequently used as reference methods in validation studies of subjective physical activity assessment.

49.7.2.1 Pedometry

Pedometers are motion sensors usually worn at the waist which respond to vertical hip accelerations during paces (Bassett et al. 1996). Pedometry output data are usually step counts. Three basic mechanisms of pedometers have been described: a spring-suspended lever arm with metal-on-metal contact, a magnetic reed proximity switch, and a piezoelectric sensor (Crouter et al. 2003; Schneider et al. 2003). As pedometers provide data on steps taken only, they only measure walking activity, but do not capture other activities such as swimming, cycling, weight lifting, or sedentary pursuits. They may underestimate steps from slower gait speeds (i.e., 0.9 meter per second) or steps with irregular gait patterns (Berlin et al. 2006). Furthermore, they do not allow an evaluation of physical activity patterns as they do not provide information relating to activity type, duration, or intensity.

Pedometers are inexpensive and light-weight portable devices that deliver reliable estimates of physical activity for use in research and clinical settings where walking is the main type of activity. They are easy to administer, which makes them a practical tool in the assessment of physical activity in large groups of almost any age. Pedometers also have the potential to promote behavior change and have therefore been used in intervention settings, for example, in the "10,000 Steps Rockhampton Project" (Queensland, Australia) (About 10,000 Steps Project 2012; http://www.10000steps.org.au). Pedometer-derived output data have been classified by Tudor-Locke and Bassett (2004) as follows: 5,000 steps/day (sedentary behavior), 5,000 to 7,499 steps/day (low activity), 7,500 to 9,999 steps/day (somewhat active), 10,000 to 12,499 steps/day (active), and 12,500 steps/day (highly active).

To reduce assessment variability and to capture lower step counts on weekends versus weekdays, pedometers should be worn for 1 week, and the average number of steps per day should be calculated from this period (Berlin et al. 2006).

Wide variation exists between pedometer models, and their costs are proportional to their accuracy. Tudor-Locke et al. (2002) reviewed the validity of pedometers compared to accelerometers, direct observation, energy expenditure, and self-report. They rated pedometry as a valid physical activity assessment method when compared with various accelerometers ($r = 0.86$) and summarized that pedometers are correlated with energy expenditure estimates from heart rate monitoring in the range of $r = 0.46–0.88$ and from indirect calorimetry in the range of $r = 0.49–0.81$ (Tudor-Locke et al. 2002).

In children, the accuracy of a pedometer was compared to accelerometer measures and direct observation. Correlations between the pedometer and the accelerometer were $r = 0.50$ during classroom activity and $r = 0.98$ during recreational activity (Kilanowski et al. 1999). Enhanced accuracy was reported for faster walking speeds compared to slower walking and running speeds (Crouter et al. 2003).

49.7.2.2 Accelerometry

Accelerometry is the most commonly used objective method to determine physical activity intensities and estimates of energy expenditure in free-living populations. Acceleration is measured by small monitor sensoring devices usually attached to the hip or lower back, but they can be also worn on the wrist, thigh, or ankle. Acceleration produced by the movement of a body segment or limb parts is measured in one (longitudinal body axis, usually vertical), two (vertical and mediolateral or vertical and anterior-posterior), or three (vertical, anterior-posterior, and mediolateral) directions (Chen and Bassett 2005), depending on the device. A piezoelectric accelerometer converts a physical force into an electric signal through the deformation of a piezoelectric element by a seismic mass, which changes the charge of the sensor causing a voltage signal that is proportional to the acceleration (Chen and Bassett 2005). The Actigraph (formerly known as the Manufacturing Technology Incorporated (MTI) and the Computer Science Applications, Inc. (CSA)) has been one of the most widely used uniaxial accelerometry device in epidemiological studies. Recent available devices are the AMP-331 (biaxial) and the Actical (multiaxial) (Crouter et al. 2006a). The use of multiple accelerometers only marginally improved the estimation of energy expenditure compared to estimates from a single accelerometer (Trost et al. 2005). However, multiaxial accelerometers are more sensitive to light-intensity activities, and they provide more valid measurements of static upper-body movement in activities such as cycling and rowing (Steele et al. 2003). The Intelligent Device for Energy Expenditure and Physical Activity (IDEEA) represents an accelerometer-based system that uses piezoelectric sensors attached to the chest, feet, and thighs (Zhang et al. 2003). Compared to commonly used accelerometers, the IDEEA uniquely enables to differentiate types of activity such as sedentary behavior (e.g., sitting, lying, or standing) and active paces (e.g., walking, cycling) and thus provides reliable estimation of energy expenditure in free-living populations (Kwon et al. 2010).

The principal output of accelerometers is counts, which are derived from the force and frequency of vertical displacement. Accelerometer counts are sampled over a preset sampling period or epoch (Ward et al. 2005). Several studies have used 1-min epochs to collect accelerometer data. However, short and frequent bouts of children may inadequately be captured using a 1-min epoch and consequently may lead to an underestimation of moderate to vigorous physical activity (Ward et al. 2005).

The conversion of physical activity counts into other measures of activity is termed value calibration (Warren et al. 2010; Bassett et al. 2012). An important methodological issue is how to translate the accelerometer signals and counts into biologically meaningful data. Usually, single linear regression equations are used to relate count values to units of energy expenditure (Crouter et al. 2006b). These estimated energy expenditure values may depend on the accuracy of the prediction equations derived from various calibration studies in laboratory or field settings, which may differ between population groups, such as children and the elderly (Hu 2008). Additionally, these regression equations allow establishing cut-points for classification of physical activity into different activity intensities, including light, moderate, and vigorous intensities (Crouter et al. 2006b).

The range of published cut-points for sedentary activities varies from <100 to <800 counts per min. For moderate-intensity activities, the range varies from 1,900 to 8,200 counts per min (Warren et al. 2010). Thus, the time spent at different intensity levels will differ considerably according to the adopted threshold. This may impact upon results regarding the relationship between physical activity intensity and health outcomes.

A comparison of calibration studies of accelerometers is difficult because of the different populations, protocols, and criterion measures used (van Cauwenberghe et al. 2011). A study by van Cauwenberghe et al. (2011) evaluated differences in physical activity cut-points by comparison of four different cut-points in a population of children (Sirard et al. 2005; Pate et al. 2006; Evenson et al. 2008; van Cauwenberghe et al. 2011). The cut-points for sedentary behavior from Pate et al. and Evenson et al. were considerably lower than the cut-points defined in the Sirard et al. (2005) study and the van Cauwenberghe et al. study. The cut-points for moderate to vigorous physical activity of the van Cauwenberghe et al. study and the Evenson et al. study were comparable, and the cut-points of Pate et al. were lower (Table 49.3). Higher moderate to vigorous activity cut-points have been established by Sirard et al. compared to the cut-points of the other studies (Table 49.3). van Cauwenberghe et al. (2011) speculated that the differences in cut-points among the studies may be primarily ascribed to limited compliance by study participants to the study protocol rather than variation in the actual activities included in the protocol itself. Figure 49.9 shows the time in moderate to vigorous physical activity according to different cut-points used in selected studies.

When using the Pate et al. cut-points for moderate to vigorous physical activity, 91% of children meet the daily moderate to vigorous physical activity recommendation, whereas none of these children meet the recommended moderate to vigorous physical activity levels according to the cut-point of Sirard et al. (Fig. 49.10). Whether the substantial differences in cut-points among these studies are of biological relevance remains to be clarified in further investigations (van Cauwenberghe et al. 2011).

Table 49.3 Physical activity intensity cut-points for preschoolers (Adapted from van Cauwenberghe et al. (2011))

	Moderate to vigorous activity	Vigorous activity
Pate et al. (2006)	420–841	≥842
Sirard et al. 2005	891–1,254	≥1,255
Evenson et al. 2008	574–1,002	≥1,003
van Cauwenberghe et al. 2011	585–880	≥881

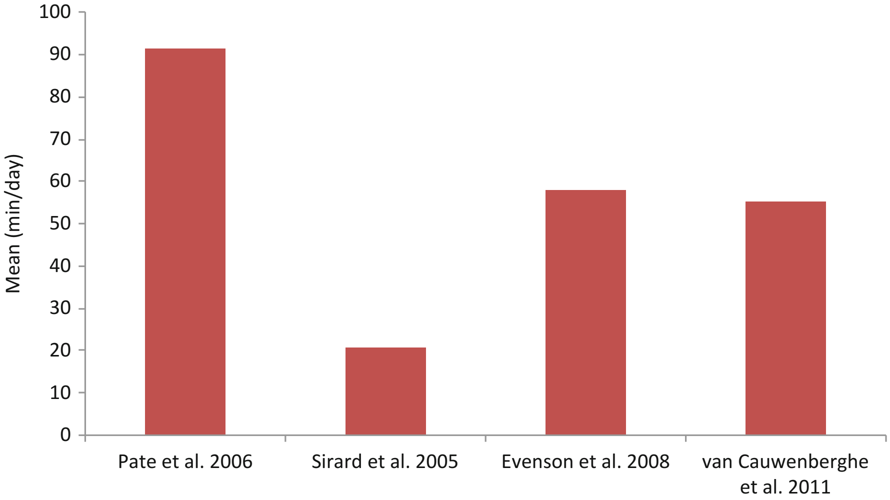


Fig. 49.9 Differences in predicted time spent in physical activity of moderate to vigorous intensity according to different cut-points used (van Cauwenberghe et al. 2011)

As accelerometers are typically worn on the waist, lower-extremity or trunk acceleration such as walking, running, and stair climbing is well captured, whereas upper-extremity movements or seated activities are inaccurately assessed (Crouter et al. 2006b). Welk et al. (2000) have shown that accelerometry underestimates the energy expenditure of household activities (Fig. 49.11) whereas it provides reasonable energy expenditure estimations of activities such as walking and jogging (Welk et al. 2000) (Fig. 49.12). Moreover, Brage et al. (2003) reported that CSA accelerometer outputs increase linearly with speed in the walking range but not in the running range, probably due to relatively constant vertical acceleration in running. Several regression equations are available that relate activity counts to energy expenditure. However, no single regression equation can accurately estimate energy expenditure across the entire range of physical activity. Crouter et al. (2006b) developed a model for the Actigraph accelerometer, which integrates a walk/run regression equation and an equation for intermittent lifestyle activities.

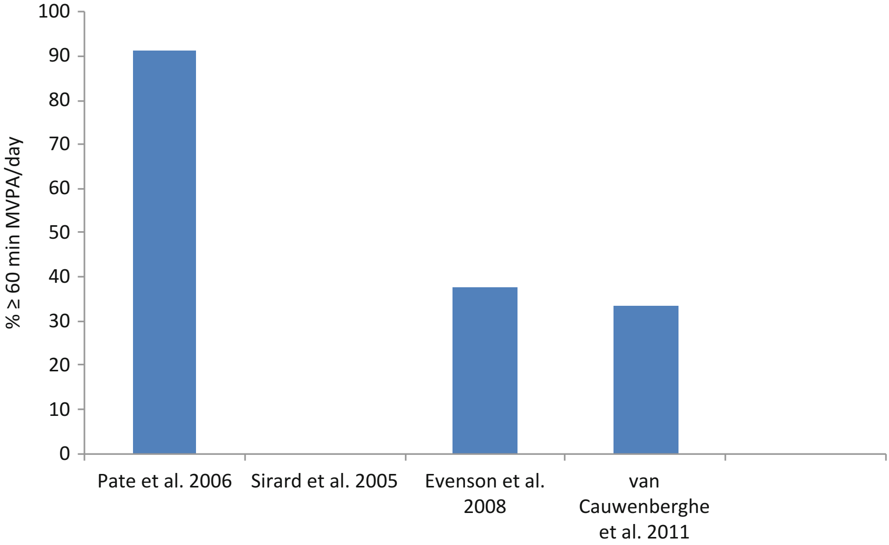


Fig. 49.10 Proportion of children and youth complying with the physical activity recommendations (≥60 min moderate to vigorous physical activity (MVPA) per day) according to different cut-point used (van Cauwenberghe et al. 2011)

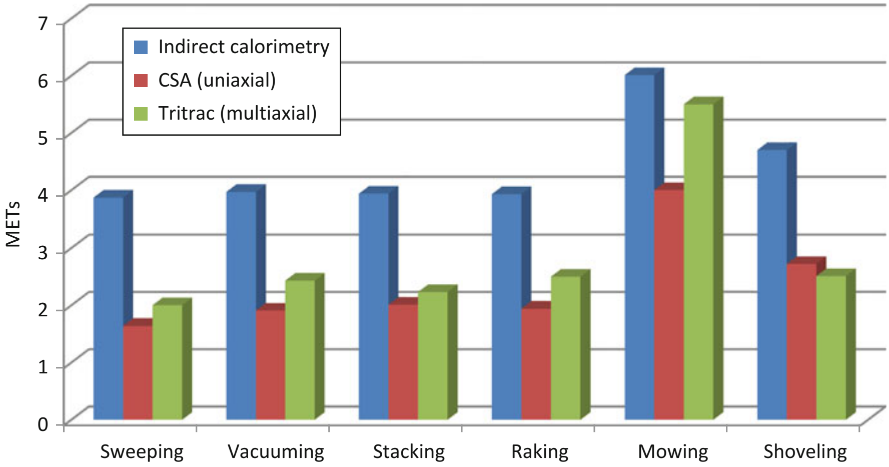


Fig. 49.11 Prediction of energy expenditure of household activities by an uniaxial and a multiaxial accelerometer (CSA and Tritrac) as compared with indirect calorimetry (Welk et al. 2000)

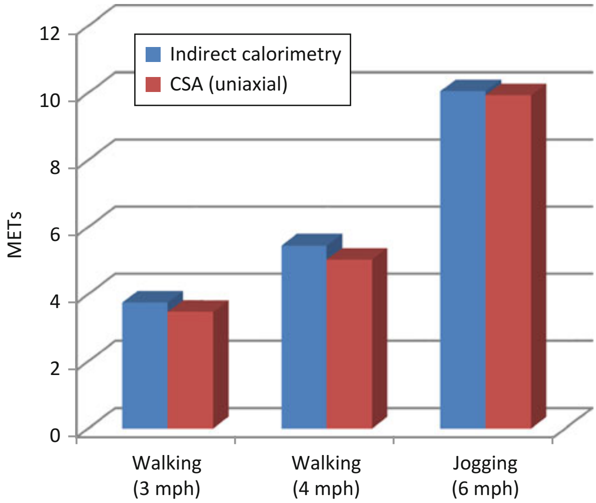


Fig. 49.12 Prediction of energy expenditure of walking and jogging by a CSA accelerometer as compared with indirect calorimetry (mph = miles per hour) (Welk et al. 2000)

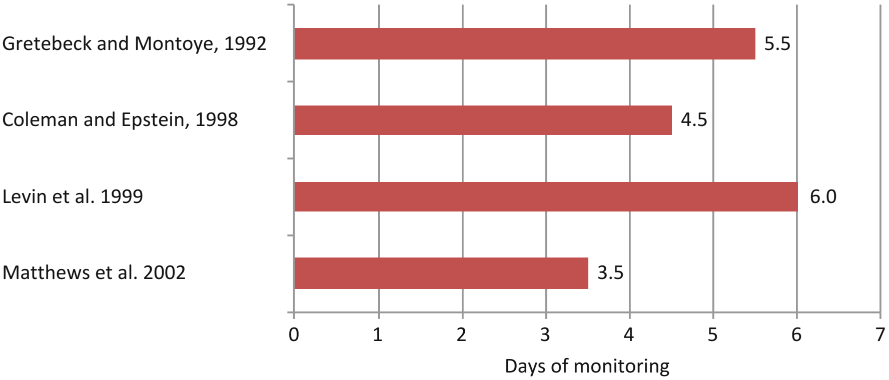


Fig. 49.13 Number of days suggested to reliably assess physical activity using an accelerometer in adults (Adapted from Trost et al. (2005))

The reliability and validity of accelerometers have been thoroughly studied. Bassett et al. (2000) assessed the validity of three accelerometers and one pedometer for estimating energy expenditure during moderate physical activity intensity against energy expenditure measured by a portable indirect calorimeter, with correlation coefficients ranging from 0.33 to 0.62.

Trost et al. (2005) summarized available research regarding the number of days required to wear an accelerometer to achieve reliable estimates of habitual physical activity. The authors suggested that adults wear an accelerometer for 3 to 5 days (Fig. 49.13). In children, the recommendations of monitoring time range from 4 to 9 days, making it difficult to draw a definitive conclusion on the precise number of days needed (Fig. 49.14). Trost et al. (2005) suggested that a 7-day monitoring period may be a reasonably sensitive period for children. An advantage

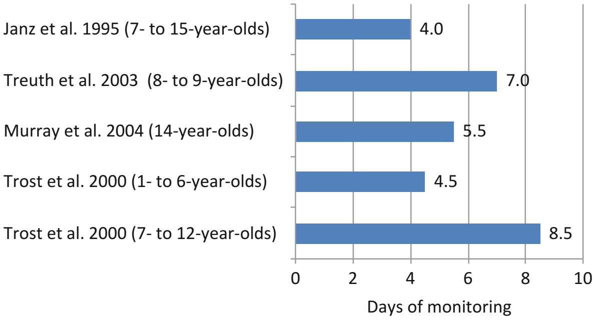


Fig. 49.14 Numbers of days suggested to reliably assess physical activity using an accelerometer in children (Adapted from Trost et al. (2005))

of accelerometers is that they provide a relatively detailed measurement of activity patterns. Moreover, their ease of administration makes them a feasible assessment tool for individuals across all age groups. The amount of data produced from accelerometers requires some skill to process and read into energy expenditure output. There is no standardization regarding the use of energy prediction equations and cutoff points for different activity intensities which vary depending on the types of calibration activities and the setting in which the calibration was performed. Imprecise estimation of energy expenditure may not necessarily be the product of imprecise motion detection but rather caused by limitations in the accuracy of the regression equations used to predict activity-related energy expenditure (Welk et al. 2000). Financial costs may prohibit the use of accelerometers for assessment of physical activity in large samples, unless accelerometer devices become less costly. Moreover, the development of accelerometers with multiple sensors may allow more detailed assessments of physical activity patterns in the future.

49.7.2.3 Heart Rate Monitoring

Heart rate monitoring represents a measure of a direct physiological response to physical activity and can be used to differentiate between activity intensities (Sirard and Pate 2001). Heart rate monitoring devices commonly comprise a chest strap and a wristwatch-type receiver for transmission of data.

Heart rate (HR) increases linearly and proportionately with the volume of oxygen consumed (VO_2) by contracting skeletal muscles in the range of moderate to vigorous physical activity. However, during light-intensity physical activity, linearity is poor since HR is influenced by several factors, such as heart rate, body temperature, muscle mass, caffeine intake, stress, type of exercise, and medication use (e.g., beta-blockers) (Acheson et al. 1980; Janz 2002; Lee et al. 2009).

Heart rate monitoring uses prediction equations that enable the evaluation of energy expenditure of physical activity in free-living populations (Vanhees et al. 2005). In order to reduce interindividual error in the prediction of energy expenditure, calibration curves on the relationship between HR and VO_2 of each individual of the study are constructed. In these calibration studies, HR and VO_2 are determined under resting conditions (lying, sitting, or standing) and during different levels of exercise (e.g., bicycle ergometry) in a laboratory setting. During field conditions,

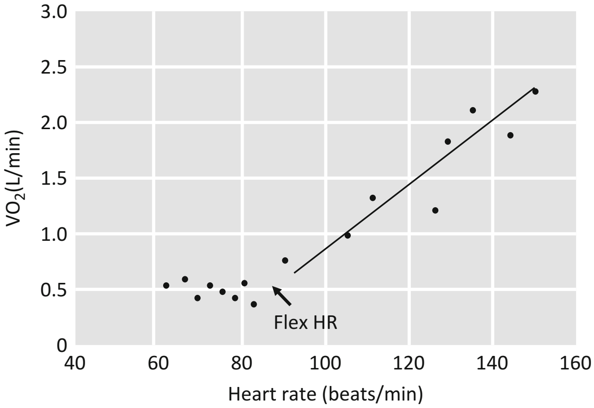


Fig. 49.15 Relation of heart rate to oxygen consumption using the flex heart rate method (Adapted from Janz (2002))

these calibration curves are then used to estimate VO_2 and energy expenditure from heart rate (Janz 2002). One potential limitation of the HR monitoring method is that it requires individual calibration.

The flex heart rate (flex HR) method represents an approach for estimation of energy expenditure from heart rate (Fig. 49.15). Flex HR is determined by measuring HR and VO_2 at rest and during exercise followed by the determination of the flex HR (Wareham et al. 1997). The flex HR is empirically determined as the mean of the highest resting value and the lowest exercise value. For HR values falling below the flex HR, energy expenditure is assumed to be equal to resting energy expenditure. For HR values above the flex HR, energy expenditure is predicted from individual HR-VO_2 calibration curves.

Energy expenditure can also be derived using the %HRR method. One approach described by Strath et al. (2000) is based on the assumption that %HRR and $\%VO_{2R}$ are approximately equal, which allows the prediction of EE through activity HR and physiological relationships. Figure 49.16 shows how energy expenditure can be calculated from HR data using this method. HR is measured under laboratory conditions. For determination of HR_{max} and $VO_{2\,max}$, prediction equations can be used. For a more precise assessment, direct measurements are required. The HR values are transformed into %HRR values using the following equation:

$$\%HRR = [activityHR - restingHR)/(estimatedHR_{max} - restingHR)] \cdot 100\%.$$

Assuming that %HRR is equal to $\%VO_{2R}$, the relative intensity of the activity can be determined. Absolute VO_2 can be calculated using $\%VO_2$ for each activity by utilizing the equation

$$\%VO_{2R} = [activityVO_2 - restingVO_2)/(estimatedVO_{2\,max} - restingVO_2)] \cdot 100\%.$$

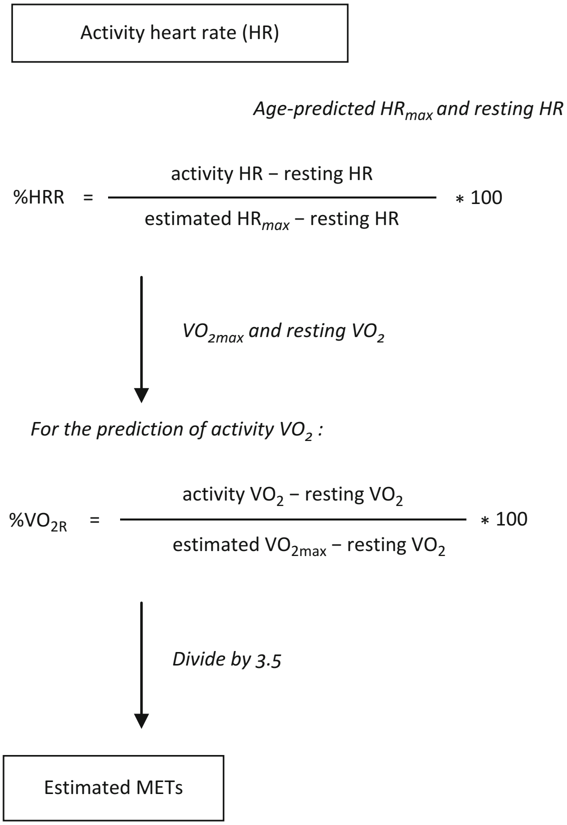

Activity heart rate (HR)
Age-predicted HR_{max} and resting HR
$\%HRR = \frac{\text{activity HR} - \text{resting HR}}{\text{estimated HR}_{max} - \text{resting HR}} * 100$
VO_{2max} and resting VO_2
For the prediction of activity VO_2 :
$\%VO_{2R} = \frac{\text{activity VO}_2 - \text{resting VO}_2}{\text{estimated VO}_{2max} - \text{resting VO}_2} * 100$
Divide by 3.5
Estimated METs

Fig. 49.16 Calculation of energy expenditure (metabolic equivalent tasks) using age-predicted percent heart rate reserve (%HRR) and estimated percent VO_2 reserve (%VO_{2R}), assuming that %HRR = VO_{2R} (modified from Strath et al. (2000))

The flex heart rate and %HRR methods have been validated by comparison with both whole-body indirect calorimetry and DLW. Validation studies have demonstrated a moderate to high degree of accuracy of both methods (Spurr et al. 1988; Livingstone et al. 1990, 1992; Strath et al. 2000). However, the reliability and validity of heart rate monitoring for measuring low-intensity physical activity are lower because of the non-linearity of the heart rate and $VO_{2\,max}$ relation during resting and light activity (Janz 2002).

Currently available heart rate monitors represent an easy and quick method for data collection, and they have a large data storage capacity (Hu 2008; Warren et al. 2010). They pose relatively low burden on the participant, although compared to accelerometers it may be more burdensome for individuals to wear chest belts. Heart rate monitoring measures physical activity of moderate to vigorous intensity well, and it shows a strong association with energy expenditure (Janz 2002).

49.7.2.4 Combination of Heart Rate Monitoring and Accelerometry

Recently, the combination of heart rate monitoring and accelerometry has been introduced for a more accurate assessment of physical activity with the idea that this method compensates for the limitations of each of its component methods. As described earlier, heart rate monitoring is less precise at low physical activity intensity levels. However, it provides reasonable accuracy at high-intensity activities. In contrast, accelerometry is more accurate at low-intensity physical activity levels and is less accurate at high-intensity levels. Moreover, physical activities such as cycling, carrying weights, and upper-body activities are not precisely assessed by accelerometers. However, they are well captured by heart rate monitoring (Brage et al. 2005; Strath et al. 2005). The measurement error from the two methods is not meaningfully correlated with one another (Brage et al. 2004).

Heart rate monitoring and accelerometry have been applied simultaneously using two separate devices. The Actiheart sensor was the first sensor that combined both methods in one device (Brage et al. 2005). Barreira et al. (2009) validated the Actiheart against an AEI Moxus metabolic cart for light, moderate, and vigorous physical activity intensities and obtained Pearson's correlation coefficient of 0.79, 0.72, and 0.80, respectively (Barreira et al. 2009). Moreover, Brage et al. (2005) examined the reliability and validity of the Actiheart sensor and concluded that the Actiheart is reliable and valid. More validation and reliability studies of combined methods are required.

49.7.2.5 Direct Observation

Direct observational procedures of free-living physical activity behaviors can provide detailed and objective information on physical activity patterns, including activity type, intensity, and duration. They also allow researchers to record factors that are related to physical activity behavior, such as behavioral determinants and environmental conditions (Trost 2007). With a growing interest in the influences of the social and physical environment on activity behavior in young people, the use of this method has increased (Trost 2007).

Direct observation has commonly been used in children as a reference method to assess physical activity. Usually, an investigator will observe a child using a specific observation system and will enter a rating of the child's physical activity level into a computer or coding form (Warren et al. 2010).

One example of an established observation system includes the System for Observing Fitness Instruction Time (SOFIT) that is based on a five category scale (lying, sitting, standing, walking, very active) and uses momentary time sampling at 10 seconds observe/record intervals (McKenzie et al. 1991). This observational system allows studying influences on diet and physical activity in children in a variety of settings. The recently developed System for Observing Play and Recreation in Communities (SOPARC) evaluates physical activity types and levels and related information in boys and girls and people attending community-based settings, such as parks and recreation areas (Brown et al. 2006).

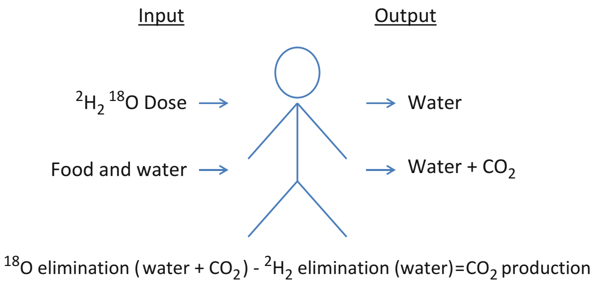

Input
Output
$^2H_2\,^{18}O$ Dose
Food and water
Water
Water + CO_2
^{18}O elimination (water + CO_2) - $^2H_2$ elimination (water)=CO_2 production

Fig. 49.17 Principle of the doubly labeled water method

The validity of the SOFIT was examined in fifth graders using a CALTRAC accelerometer ($r = 0.74$) (Sharma et al. 2011) and in elementary and middle school children using heart rate monitoring ($r = 0.80–0.91$) (Rowe et al. 1997). Interobserver reliability ratings of the SOFIT were >90% for estimates of energy expenditure, total energy expenditure, and time spent in moderate to vigorous activity (McKenzie et al. 2000).

The main strength of the direct observation method is that it provides broad information about physical activity behavior and related factors. However, direct observation does not allow assessment of activity intensity or energy expenditure. Direct observation can only be undertaken in controlled settings or small samples, and it requires considerable time and well-trained observers. In addition, the presence of an observer may lead to a reactivity effect, that is, to a change in habitual behavior (Trost 2007; Warren et al. 2010).

49.7.2.6 Doubly Labeled Water

DLW represents a method by which orally administered doses of the stable isotope tracers ^2H and ^{18}O are used to calculate TEE through carbon dioxide (CO_2) production in the human body. The study participant drinks a defined amount of water labeled with ^2H and ^{18}O and provides urine samples for the next 2 weeks. The ^2H will be eliminated from the body as water, whereas the ^{18}O will be eliminated from the body as water and CO_2 (Fig. 49.17). The difference between the elimination rates is proportional to CO_2 production. The CO_2 production rate is converted to energy expenditure using additional information on the respiratory quotient of food ingested (Starling 2002).

DLW is often considered the gold standard among physical activity assessment tools, and it has been widely used as a criterion method to validate other methods of assessment of physical activity, such as questionnaires, diaries, logs, and accelerometers. The method has been used in adults, children, and infants.

Validation studies of the DLW method against respiratory gas exchange suggest that the DLW method overestimates daily energy expenditure by 4% to 8% as compared with respiratory gas exchange (Schoeller and Webb 1984; Schoeller et al. 1986).

Strengths of the DLW method are that it precisely measures energy expenditure in adults and children in free-living conditions, and it also does not induce any change in physical activity behavior. However, one limitation of this method is that the production and analysis of the isotopes ^{2}H and ^{18}O are expensive and therefore not suitable for large-scale studies. Also, the DLW method does not provide information on specific activities performed. Furthermore, it does not allow distinguishing between RMR, AEE, and TEF. The combined DLW and indirect calorimetry application can overcome this limitation and may provide a robust measure of physical activity energy expenditure (Vanhees et al. 2005; Lee et al. 2009).

49.7.2.7 Indirect Calorimetry

Indirect calorimetry represents a physical activity measurement method that allows an estimation of energy expenditure. This method is based on measurements of oxygen uptake and carbon dioxide production with which energy expenditure can be established by equations (Starling 2002). The individual wears a canopy in the laboratory or a mask under field conditions for more extended measurement times, the investigation is performed in a metabolic chamber. The principle of this method is that controlled room air with known concentrations of O_2 and CO_2 is introduced to the plastic canopy or metabolic chamber and concentrations of O_2 and CO_2 are measured as the air leaves the system (Weir 1949; Vanhees et al. 2005). Energy expenditure can be calculated by entering the volumes of oxygen consumed and CO_2 produced into Weir's equation for calculating resting energy expenditure (REE) (Weir 1949):

$$\mathrm{REE} = [\mathrm{VO}_2(3.94) + \mathrm{VCO}_2(1.11)]\ 1{,}440\,\mathrm{min/day}.$$

Like DLW, indirect calorimetry is relatively expensive and burdensome for the participant and is therefore not feasible for large samples in epidemiological studies. Because of its high accuracy, indirect calorimetry is a widely used criterion measure for validating other physical assessment methods (Vanhees et al. 2005; Hu 2008). New portable devices for measuring RMR may provide an enhanced application of this method in the future (Nieman et al. 2005).

49.7.3 New Technologies in Physical Activity Assessment

Computers, other electronic media, and the internet represent advanced tools that are increasingly used to assess physical activity. Recently, mobile phones, smart phones, Geographic Positioning System/Geographic Information System (GPS/GIS) technology, and speech recognition have been introduced to physical activity promotion in free-living settings. Mobile phones and smart phones represent convenient devices to collect data on a large scale. Some mobile phones include integrated pedometers to monitor physical activity, although limited battery life still represents a disadvantage of these devices. Recently, speech recognition and text classification have been incorporated into physical activity assessment instruments. In speech recognition, recorded physical activity is transmitted into text via electronic circuits. Text classification allows documents to be assigned to predetermined categories according to a specific topic or function. Emerging technologies, such as GPS and GIS, represent tools that enable an accurate tracking of the interaction between people's physical activity and the surrounding environment, such as the route and temporal pattern of a walk or other movement (President's Council on Physical Fitness and Sports 2008). Recently, cell phone-based and other electronic-based diaries have been introduced (Zhu et al. 2006; Sternfeld and Dugan 2011). Zhu et al. (2006) developed an E-diary system by using voice recognition and text classification technology. Participants are reminded by a signal from a watch to record their past physical activity. They talk into a small digital recorder and record their physical activity during the past hour. The recorded voice is then converted to a text form and assigned to an activity category (e.g., "I walked the dog for 15 minutes" is classified into the category "walking") with a validity of 87% (Zhu et al. 2006).

49.7.4 Selection of a Physical Activity Assessment Method

The choice of an appropriate physical activity assessment method usually represents a trade-off between validity and feasibility but should primarily be motivated by the research question (Warren et al. 2010). Considerations that additionally impact on the selection of the method include the population studied, costs, participant burden, the capacity to perform the appropriate data handling and analysis, and experience in the assessment of physical activity (Warren et al. 2010). The study design and the physical activity outcome indented to be measured, which are central to the research question, may narrow the choice of potential physical activity assessment instruments. For example, if a ranking of individuals according to levels of activity for assessing associations in an observational study is required, one may choose between self-reported and accelerometry methods. Both methods provide physical activity outcome data and can be used at the population level, but the accelerometry method is usually restricted to a smaller sample size than the self-reported method. If information only on walking is required, then pedometry might be an appropriate method. The combination of two methods might compensate for the limitations of each method. Figure 49.18 shows available tools to be chosen to assess physical activity with reference to specific study designs.

Particular considerations are necessary when conducting studies in specific populations. For example, children's physical activity patterns consisting of short and frequent bouts and cognitive immaturation make it more challenging to select an appropriate assessment method. Direct observation is regarded as an accurate measurement tool to assess children's physical activity behavior. However, its limiting factors (costs, time, and data processing) may question the suitability of this method in certain situations. For children less than 10 years of age, a proxy report (e.g., by parents) may be the method of choice, although the limited validity of proxy reports should be kept in mind (Dollman et al. 2009). Previous day recalls may be appropriate for 10- to 11-year-olds, and activity diaries may be used in adolescents (Warren et al. 2010). Some researchers suggest using accelerometers or

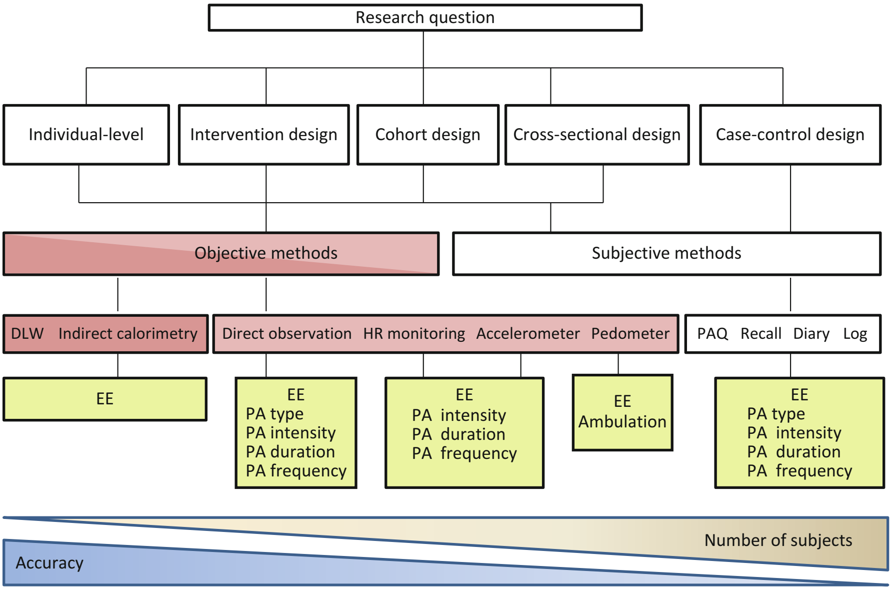


Fig. 49.18 Physical activity assessment methods according to different study designs; *DLW* doubly labeled water, *EE* energy expenditure, *PAQ* physical activity questionnaire, *PA* physical activity, *HR* heart rate (modified from Dollman et al. (2009) and Mueller et al. (2010))

a combination of accelerometers with self-reported methods for accurate assessment of children's physical activity patterns (Harris et al. 2009). To date, there is no unified concept regarding the optimal instrument to assess physical activity in youth.

Assessment of physical activity in the elderly may differ from that among young to middle-aged adults because the main activity in the elderly comprises walking and sedentary behavior. Motion sensors have been suggested as an appropriate tool, but they do not provide detail on physical activity type. Memory of past physical activity is also a concern in older people. Here, a combined approach of objective and self-report measures may allow a more accurate characterization of physical activity behavior (Harris et al. 2009).

49.8 Epidemiological Study Designs in Physical Activity Research

Epidemiological research of physical activity encompasses various study designs, including experimental studies such as randomized controlled trials (RCTs) and observational studies such as cohort studies, case-control studies, and cross-sectional studies. RCTs are generally considered the gold standard of study designs as they are most likely to yield unbiased results (Victora et al. 2004). Due to issues related to compliance, sample size, costs, and ethical considerations, the RCT design has rarely been used in physical activity epidemiological research (Caspersen 1989; Lee et al. 2009). By comparison, observational data from cohort and case-control studies are the main sources of evidence on the relationship between physical activity and risk of diseases (McTiernan 2011). This section describes the main study designs used in physical activity epidemiology including their strengths and limitations.

49.8.1 Cross-Sectional Studies of Physical Activity

The cross-sectional study design is used to examine correlations between engagement in physical activity and risk factors of diseases within populations at a certain point in time. Results from those studies may generate hypotheses and may identify potential mediators that can be targeted in intervention studies (Sallis et al. 2000; Bauman et al. 2002). Cross-sectional survey data are used by governments and institutions to provide descriptive statistics of physical activity at the population level (Garber et al. 2011). A cross-sectional study has the advantage over other studies that it is relatively easy and inexpensive to conduct by using routinely collected data (Mann 2003). However, the cross-sectional study design does not allow establishing temporal relationships between levels of physical activity and disease outcomes, as both variables are typically measured simultaneously, and thus, causality cannot be inferred from cross-sectional physical activity data. Moreover, linking changes of physical activity over time to disease outcomes may be crucial in establishing physical activity and disease relations. The NHANES was initially designed as a cross-sectional study, and it subsequently developed into a follow-up study when participants of the first survey were reexamined 10 years later. It provides both subjectively assessed physical activity by self-reported questionnaires and objectively measured physical activity via the accelerometry method. As mentioned above, NHANES also serves as a well-established surveillance system to provide nationwide descriptive data on physical activity and other health and risk factors of the US population (Centers for Disease Control and Prevention 2012).

49.8.2 Case-Control Studies of Physical Activity

Case-control studies have frequently been conducted to explore associations of physical activity with chronic diseases. For example, considerable evidence from case-control studies exists for an inverse association between physical activity and cancers of the colon, breast, and prostate (Colditz et al. 1997; Friedenreich 2001).

Cases with disease and control individuals who do not have that disease and represent the population that gave rise to the cases are asked to recall their past physical activity participation (Caspersen 1989). Compared with their control counterparts, cases may differentially recall or report their usual levels of physical activity (Caspersen 1989). Such potential differences in recall or reporting can lead to a biased estimate of the true relation of physical activity to the health outcome. A further methodological concern of case-control studies is selection bias.

For example, controls that are willing to participate in a case-control study are frequently more health conscious than subjects who decline to participate. Such health consciousness may be linked to increased physical activity among controls who choose to participate compared to subjects who decline to take part in a study, leading to potential selection bias.

The case-control study design in physical activity is best illustrated by an example. From 1995 to 1997, a population-based case-control study of 1,233 incident breast cancer cases and 1,237 controls was conducted in Alberta, Canada, that investigated the effect of lifetime physical activity patterns on breast cancer risk (Friedenreich et al. 2001). Lifetime physical activity was assessed by interviewer-administered questionnaires. No association between physical activity and breast cancer was found for premenopausal women. For postmenopausal women in the highest versus the lowest quartile of lifetime total physical activity, the adjusted odds ratio (*OR*) was 0.70 (95% confidence interval (CI) = 0.52, 0.94) and the strongest risk reductions were observed for household and occupational physical activity. The authors discussed the possibility of selection bias because of the modest response proportion among the controls (56.5%). Misclassification of physical activity through recall bias may also have been an issue in this study since respondents were asked to report detailed lifetime physical activity patterns retrospectively, and the detail of such reporting may have differed between cases and controls (Friedenreich et al. 2001). A classic retrospective case-control study design only allows ascertainment of physical activity behavior patterns using subjective methods such as PAQs or other recall instruments. Alternatively, one might conduct a nested case-control study by drawing cases and controls from a cohort study where objective measurements have been performed at baseline. This also overcomes the problem of selection bias.

49.8.3 Prospective Cohort Studies of Physical Activity

Observational studies of physical activity provide most of the evidence-based data available that support a beneficial effect of physical activity on the risks of morbidity and mortality (Garber et al. 2011). Particularly, data from prospective cohort studies revealed that increased physical activity participation may reduce the risk of CVD, cancer, and type 2 diabetes (Chomistek et al. 2012). Many studies have addressed the dose-response relationship of physical activity on different health outcomes for which a cohort study provides a superior study design than an RCT. Given a limited number of predetermined intervention arms, an RCT might not be able to capture the variety of physical activity behaviors including intensity, frequency, and duration in a real-life setting (Lee et al. 2009).

In cohort studies, participants are classified according to their physical activity status and are followed over time until the occurrence of disease. The possibility of recall bias is reduced as physical activity habits are assessed before the occurrence of the outcome (Hu 2008). However, this approach also faces methodological obstacles (U.S. Department of Health and Human Services 1998). Cohort studies of chronic conditions such as coronary events or cancer are expensive, as they require large sample sizes and long follow-up periods, and they also may be affected by selection bias when individuals are lost to follow-up. Furthermore, cohort studies are prone to reverse causation which can occur when a severe disease causes a reduction in physical activity levels rather than vice versa, which may lead to spurious results (Rockhill et al. 2001).

In prospective cohort studies, the most common methods used to query physical activity patterns are subjective methods, such as PAQs, or recall methods, but objective methods such as accelerometers have also been used.

The Nurses' Health Study represents an example of a prospective cohort study design including 121,701 female registered nurses aged 30 to 55 years who in 1976 returned a mailed questionnaire about their medical history and lifestyle. Since study enrollment, information about their medical history and lifestyle has been assessed biennially by self-administered follow-up questionnaires. In 1980, women were asked about their recreational physical activity habits, including the average number of hours spent each week during the previous year on activities such as heavy gardening, vigorous sport, running, walking or striding, bicycling, or heavy housework. In the 1982 questionnaire, the question was slightly modified ("For how many hours per week, on average, do you engage in activity strenuous enough to build up a sweat?"). In later years (from 1986 onwards), the physical activity questionnaires inquired about more detailed physical activity patterns. The Nurses' Health Study provides the possibility to perform long-term follow-up, including repeated measurement of physical activity which enables observation of changes of physical activity patterns over the life course. The study has investigated a large number of physical activity and health outcome relations, such as CVD, hip fractures, various cancers, and type 2 diabetes (Lee et al. 2009). Other prospective cohort studies that have addressed physical activity research questions include the National Institutes of Health (NIH)-AARP Diet and Health Study (Leitzmann et al. 2007), the American Cancer Society Cancer Prevention Study (Patel et al. 2008), the California Teachers Study (Mai et al. 2007), the Whitehall Studies (Sabia et al. 2012), and numerous other studies.

49.8.4 Randomized Controlled Trials of Physical Activity

While observational studies of physical activity primarily consider physical activity behavior as an exposure, RCTs have mostly focused on exercise. The RCT study design supports investigations on effects of exercise training on physical fitness and biomarkers of chronic disease or other intermediate factors (Garber et al. 2011). Thus, RCTs allow an unraveling of biological mechanisms to gain insight into etiological pathways linking physical activity to diseases which cannot be captured by observational studies. Compared to observational studies, findings on physical activity-related health effects resulting from RCTs are less likely to be confounded. For example, body mass index (BMI) is considered a major factor that influences physical activity and health outcome associations, and RCTs may only be minimally fraught with confounding by BMI. However, RCTs may not always be feasible to investigate CVD, cancer, or mortality as those endpoints require large sample sizes and/or long follow-up periods. Besides financial and ethical constraints, high dropout rates raise the possibility of selection bias, and non-adherence of the participants to the study protocol represents a potential concern (Caspersen 1989).

The multilateral nature of physical activity behavior includes various types and components, such as duration, intensity, and frequency, which may not be adequately captured in an RCT with only a limited number of intervention arms. As shown by the short time window in which a large number of findings in physical activity research have been produced in the past years, ongoing RCTs may not be able to follow the dynamic feature of physical activity research as their finite protocol does not allow the flexibility needed to tackle emerging issues and new research questions (Lee et al. 2009).

49.9 Methodological Considerations of Studies on Physical Activity and Disease

Each physical activity assessment method is characterized by specific strengths and limitations, and there is no single comprehensive measurement approach that can rule out all potential challenges. Methodological considerations include the precise assessment of physical activity, control for potential confounding, detection and reporting of effect modification, and a reasonable interpretation of the relationship between physical activity and disease outcome (Courneya and Friedenreich 2011). This section reviews methodological considerations that need to be taken into account when conducting studies and analyzing data on physical activity in relation to disease outcomes.

49.9.1 Measurement of Exposure

Discrepancies between studies examining the association between physical activity and a specific health outcome may arise from misconception and misclassification of the term physical activity. This may lead to inconsistency in physical activity assessment and in heterogeneity of the results. Also, systematic error in measurement of physical activity or covariates may lead to underestimation or overestimation of the association with a disease outcome (Schatzkin et al. 2009). In particular, estimating energy expenditure from self-reported PAQs is vulnerable to measurement error. For example, overreporting of the level of physical activity may lead to an overestimation of the true effect of physical activity on disease outcome. Compared to self-reported physical activity assessment, a stronger relationship between objectively measured physical fitness and all-cause mortality has been observed (Lee et al. 2011). Combining subjective and objective measures of physical activity may alleviate the limitations in estimation of energy expenditure by subjective measures alone.

Physical activity is an exposure variable that does not comprise a single factor, but rather includes a cluster of components, including intensity, frequency, and duration of recreational, occupational, transportation, and household physical activity (Caspersen et al. 1985). A greater reduction in breast cancer risk has been observed with vigorous compared to moderate-intensity activity (Friedenreich and Cust 2008). Also, a breast cancer risk gradient according to physical activity domain has been reported, with the greatest risk reduction in breast cancer observed for recreational activity (20% risk reduction), followed by walking/cycling for transportation (14% risk reduction), household (14% risk reduction), and occupational activity (13% risk reduction) (Friedenreich and Cust 2008). Thus, it is important to distinguish between physical activity components and domains when exploring the effect of physical activity on different health outcomes of interest.

Physical activity intensity has been categorized into levels of light, moderate, and vigorous intensity by using MET values as cut-points. MET values vary between age groups, and they also differ between obese and non-obese persons (Garber et al. 2011). Furthermore, the capacity for $VO_{2\,\max}$ varies by fitness status. Thus, when using intensity categories, one should bear in mind that MET values for lean persons might not be appropriate for the elderly or for obese persons. For example, 3 METs might represent light-intensity activity for a young and fit person but moderate or vigorous intensity for an older person.

Most studies linking physical activity to subsequent health outcomes have used baseline physical activity as an exposure. Repeated measurements of physical activity in prospective cohort studies may reduce misclassification of physical activity that arises from intraperson variation (Hu 2008). For example, the Nurses' Health Study found a stronger association between physical activity and the risk of cholecystectomy in a 10-year follow-up when using the cumulative average physical activity generated over subsequent follow-up periods than by using baseline physical activity measurement or the most recent assessment (Leitzmann et al. 1999). Moreover, repeated measurements allow teasing out different patterns and changes of physical activity over time which cannot be captured by a single measurement. It has been proposed that physical activity performed during different time periods in life, such as early, mid-, or later life, is important for reduction in risk of some cancers (Friedenreich and Cust 2008). These observations suggest that inconsistent findings in epidemiological studies of physical activity may be ascribed to differences in the time periods for which physical activity was assessed.

Studies of physical activity and disease or mortality frequently include analyses of dose-effect associations. Indeed, dose-effect relations with physical activity have been reported for a large number of diseases (Kesaniemi et al. 2001). Notwithstanding, misclassification of physical activity level may potentially occur when assigning divergent cut-points for dose-response analyses of continuous measures of physical activity (Lamonte and Ainsworth 2001). For example, a spurious result might occur when cut-points are chosen that maximize the desired effect of physical activity on a disease outcome rather than cut-points that are based on physiological threshold values.

Measurement error correction methods can be used to correct for both random and systematic errors but have rarely been used in epidemiological studies of physical activity. One study examined the effect of measurement error in estimating questionnaire-based physical activity by using data from a subsample of participants who completed four 7-day diaries of physical activity over a 1-year period (Leitzmann et al. 1999). That study reported a relative risk of cholecystectomy of 0.85 (95% CI = 0.80–0.90) associated with each 25-MET per week increase in physical activity that was reduced to 0.58 (95% CI = 0.43–0.77) after adjustment for errors in measuring physical activity.

49.9.2 Confounding

Observational studies of physical activity that do not adequately control for potential confounding factors may yield bias associations between physical activity levels and health outcome. Age, BMI, dietary and alcohol intake, smoking habits, and other health behaviors have been proposed to affect the association of physical activity with disease outcomes (Friedenreich 2001). Statistical adjustment for such potential confounding factors is necessary to evaluate the independent effect of physical activity on disease outcome. For example, inconsistent findings on the relationship between physical activity levels and risk of lung cancer have been documented. One study observed an inverse association between levels of physical activity and lung cancer risk which progressively became attenuated with increasing control for smoking intensity (Leitzmann et al. 2009). Moreover, physical activity was inversely related to lung cancer among current and former smokers but was unrelated to lung cancer risk among never smokers. Because smoking is associated with both levels of physical activity and lung cancer risk and is imperfectly measured, the authors assumed residual confounding by smoking as an alternative explanation for the apparently protective effect of physical activity on lung cancer risk among smokers. They argued that if there was a causal relationship between physical activity and lung cancer, an inverse association would be observed among both smokers and non-smokers. In studies of physical activity, confounding may also occur within different components of physical activity, each of which contributes to the volume of energy expended. For example, a relationship between a specific intensity of physical activity and the risk of disease may be due to an increased volume of the energy expended rather than to the intensity of the activity itself (Lee et al. 2009). In addition to adjustment for the volume of energy expended, adjustments should be made for the intensities of the other activities performed. For example, persons who engage in sports (vigorous intensity) may also more likely walk or use a bicycle for transportation (moderate intensity), the latter of which may constitute a confounding effect in an analysis of the effect of sports activities (Lee et al. 2009).

49.9.3 Effect Modification

Effect modification or interaction occurs when a third factor modifies the effect of physical activity on a health outcome. In physical activity studies, effect modification has been less studied, and little is known about dietary and other lifestyle factors modifying physical activity and disease relations (Slattery and Potter 2002). Sometimes, effects of physical activity vary among subgroups, depending on the characteristics of the population and other factors. One should bear in mind that interactions may also give rise to significant associations by chance (Willett 1998). Low sample size frequently poses a limitation for detecting effect modification and performing subgroup analyses. However, an understanding of modifying effects of physical activity can be useful to identify and explain differences in associations between physical activity and disease outcomes. Moreover, knowledge of modifying effects of specific factors on physical activity and disease associations may provide a better understanding of potentially relevant biological mechanisms (Slattery and Potter 2002).

For example, Patel et al. (2008) examined the role of BMI as a potential effect modifier of the relation of physical activity to endometrial cancer risk among women aged 50 to 74 years (Patel et al. 2008). They observed that increased physical activity was related to a decreased endometrial cancer risk and that BMI significantly modified that association. Specifically, an inverse relationship between physical activity and endometrial cancer risk was seen only among overweight or obese women (relative risk (RR) = 0.59; 95% CI = 0.42–0.83) but not in normal-weight women (RR = 1.01; 95% CI = 0.69–1.48) (Patel et al. 2008). Such data help increase our knowledge regarding the etiological pathways that potentially link increased physical activity with decreased endometrial cancer risk. Specifically, unopposed estrogen is a major determinant of endometrial cancer, and physical activity may decrease estrogen levels in postmenopausal women by reducing peripheral adipose tissue, which is the main source of estrogen production in postmenopausal women. Moreover, physical activity may affect endometrial cancer risk by reducing hyperinsulinemia, even in the absence of weight loss (Kaaks et al. 2002; Patel et al. 2008).

49.9.4 Mediation

A mediator is a variable that is affected by the exposure and lies on the causal pathway between exposure and disease outcome. In studies of physical activity and risks of cardiovascular disease (CVD) or diabetes, the variables blood pressure, high-density lipoprotein (HDL) cholesterol, and inflammatory markers have been suggested as intermediates on the causal pathway (Kesaniemi et al. 2001; Pischon et al. 2003). The Honolulu Heart Program investigated the association between physical activity and the incidence of coronary heart disease (CHD) and mortality during a 23-year follow-up in a cohort of 8,006 middle-aged men of Japanese ancestry (Rodriguez et al. 1994). They found a reduction in CHD morbidity and mortality with increased physical activity when adjusting for age only. When additionally including other risk factors for CHD in the multivariate model, such as hypertension, BMI, serum cholesterol, and diabetes, the physical activity effect estimate was attenuated and was no longer statistically significant, suggesting that those CVD risk factors may mediate the effect of physical activity on CHD. Controlling for intermediate variables that are on the causal pathway of disease can lead to statistical overcontrol and underestimation of the true effect of physical activity on disease outcome.

Sallis et al. (2000) identified several important correlates of physical activity, including social, psychological, demographic, biological, and physical environment factors. In future controlled trials of physical activity and disease outcomes, such correlates of physical activity may be used as mediating or moderating variables and may contribute to the identification of biological mechanisms (Bauman et al. 2002).

49.9.5 Reverse Causation

Epidemiological studies investigating the association of physical activity with a disease outcome are potentially prone to reverse causation. For example, an undiagnosed preclinical disease may cause a decline in physical activity, leading to a spurious relationship between physical activity behavior and the outcome. Rockhill et al. (2001) examined the association between recreational physical activity and mortality in middle-aged and older women in the Nurses' Health Study. They found a stronger inverse association between physical activity and respiratory death and death from other causes (e.g., dementia, chronic obstructive pulmonary disease, or cirrhosis of the liver) than deaths from cancer, CVD, and diabetes. The authors reasoned that part of physical activity and mortality relationship may be spurious and ascribed to reverse causation because women who died from respiratory disease or dementia were more likely to report major activity limitations before death than women who died from cancer or CVD (Rockhill et al. 2001). Analytical approaches to avoid reverse causation are as follows: (1) exclusion of individuals with chronic diseases such as cancer or CVD before or at study baseline, who may have become physically inactive because of health problems; stratification by the presence or absence of chronic diseases at baseline might also be possible; (2) control for potential confounding variables; if repeated measures are done, controlling for updated variables or changes in variables during follow-up is desirable and (3) ceasing the updating of physical activity data when an individual is diagnosed with chronic conditions such as CVD or cancer; also, conducting sensitivity analysis by excluding those with early deaths after diagnosis of disease (e.g., patients with CVD or cancer who died during the early years of follow-up) (Rockhill et al. 2001; Hu 2008).

49.9.6 Intra-Individual Variation: Day-to-Day and Seasonal Variability in Physical Activity Behavior

Physical activity levels among free-living humans vary from day to day and may depend on season, population characteristics, the physical activity component measured, and the days sampled (Matthews et al. 2001a; Ridley et al. 2009). Matthews et al. (2001a) examined sources of variance in physical activity levels

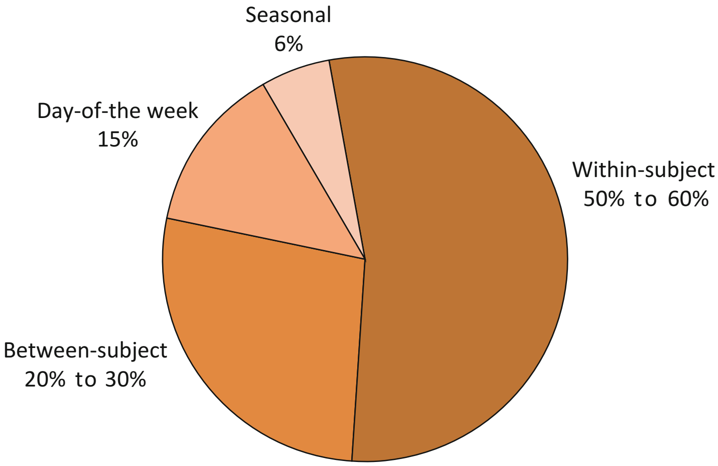


Fig. 49.19 Sources of day-to-day variation in total physical activity (Matthews et al. 2001a)

among adults in the Seasonal Variation of Blood Cholesterol (SEASON) study. The greatest source of variance in total activity was within-subject variance (50% to 60%), followed by between-subject variance (20% to 30%), day-of-the-week variance (15%), and seasonal effects (6%) (Matthews et al. 2001a) (Fig. 49.19).

Seasonal variation in physical activity is related to temperature, precipitation, and daylight (Matthews et al. 2001b; Merrill et al. 2005). In several studies, higher physical activity participation was observed in summer months than in winter months (Matthews et al. 2001b; Merchant et al. 2007). McCormack et al. (2010) investigated seasonal variability of different types and intensities of physical activity in adults and found that compared with winter, participation in walking for recreation was increased in the summer, autumn, and spring and that walking for transportation was increased in autumn. Moderate physical activity was also more likely in the summer. In addition, they observed age- and sex-dependent variation in vigorous physical activity participation (McCormack et al. 2010). Seasonal variability in physical activity patterns has also been shown to vary by BMI and type of physical activity. Matthews et al. (2001b) reported a smaller increase in summer recreational activity in obese or sedentary women compared to lean and active women. Sedentary men and women showed less seasonal variation in recreational activity than men and women who engaged in exercise and sports (Matthews et al. 2001b).

Nilsson et al. (2009) studied 9-year-old children and 15-year-old adolescents in four European countries (Denmark, Portugal, Estonia, and Norway) and found between- and within-day differences in overall physical activity levels, time spent in moderate to vigorous activity, and sedentary pursuits, which differed by sex, age, and geographical location. Specifically, engagement in 60 minutes of moderate- to vigorous-intensity activity and time spent in sedentary behaviors tended to be higher during weekdays than weekend days. Moreover, school-time and leisure-time activities varied between countries (Nilsson et al. 2009). Figure 49.20 shows

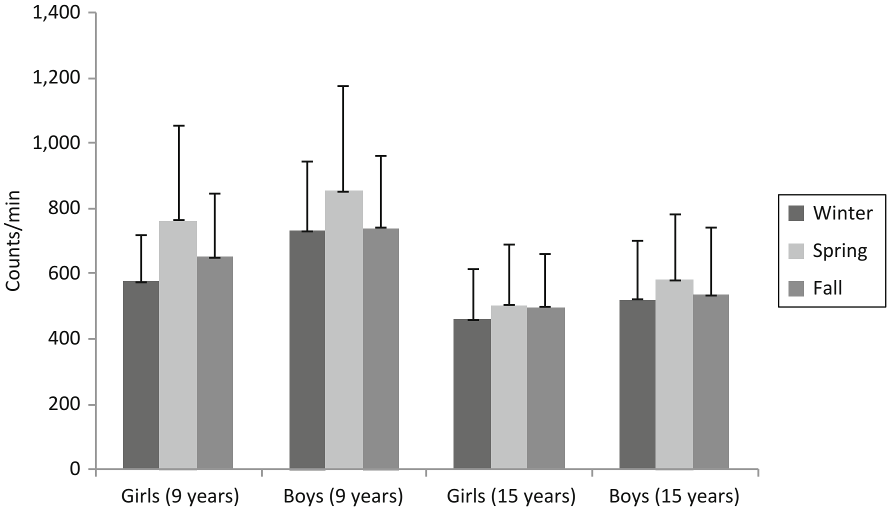


Fig. 49.20 Physical activity levels among children and youth stratified by age, sex, and season (Kolle et al. 2009)

physical activity patterns among 9- and 15-year-olds in Norway, which is a climatically diverse country (Kolle et al. 2009). Nine-year-old children showed significantly higher physical activity levels in the spring than in the winter or fall. By comparison, no differences in physical activity were observed among the 15-year-olds (Fig. 49.20) (Kolle et al. 2009).

Those data emphasize the importance of taking day-to-day variation in physical activity into account because errors may result from intra-individual variability in physical activity. Before conducting a study, an assessment of the number of days required to reliably characterize physical activity behaviors is required. An understanding of the nature of the variability in physical activity pattern in the population under study is necessary for an adequate design and analysis of the study and an appropriate interpretation of physical activity results (Matthews et al. 2001a). Measurement of physical activity at a single point in time is insufficient to capture variation in physical activity patterns. Physical activity assessment over 1 year captures seasonal variability in a population (Matthews et al. 2001a). Paying attention to day-to-day variation in physical activity behavior of individuals including seasonal variability may be important when designing intervention programs.

49.9.7 Causal Inference

The most widely cited epidemiological criteria of causality are those of Sir Austin Bradford Hill (1897–1991), who was the first to define nine criteria for judging whether an association between an exposure and an outcome is causal: consistency, strength, specificity, dose-response relationship, temporality, biological plausibility, coherence, and experiment (Hill 1965). Susser (1991) and the 1964 Surgeon General's Report (1964) used similar criteria to consider causal relationships, and all those criteria have temporality, strength, consistency, and specificity in common (Susser 1991; Department of Health, Education and Welfare 1964; Parascandola et al. 2006). However, these criteria and their implementation in scientific causality judgment have been widely debated. Hill's criterion of consistency encompasses the replication of findings involving different persons, places, circumstances, and time (Hill 1965). Given the multidimensionality and complexity of physical activity research, consistency might not entirely allow judging the causality of an association. From a logical perspective, a strong association between physical activity and a disease outcome might be more likely to suggest causality than a weak or modest association (Hoefler 2005). However, even weak physical activity associations might have biological meaning. In general, associations with disease outcomes are weaker with physical activity than they are with clinical parameters. Nevertheless, physical activity may represent a causal component of disease development. Causation of most diseases relies on the combination of multiple factors, such as dietary, lifestyle, genetic, psychological, and environmental factors, and the causal effect may be the sum of several small effects. Rothman and Greenland (2005) stated that strength does not represent a biological mechanism and may depend on the occurrence of other causes. In physical activity research, the strength of a physical activity effect on the occurrence of a disease may change over time as the prevalence of the determinants of physical activity changes even though the causal mechanisms leading to disease might remain unchanged. Temporality has been the most accepted criterion, assuming that a putative cause precedes the effect. In studies of physical activity in relation to disease incidence, it might be difficult to establish the temporal sequence between physical activity and the disease. For example, temporal sequence can be blurred when the disease has a long induction period (Rothman and Greenland 2005). In physical activity-disease relations in particular, it can be challenging to establish a temporal sequence as physical activity and physical condition are tightly linked, and disease occurrence may lead to reduced physical activity before diagnosis was made. Finally, it was Hill himself who stated: "none of my nine viewpoints can bring indisputable evidence for or against the cause and effect hypothesis and none can be required as a sine qua non" (Hill 1965).

49.10 Areas of Potential Future Development in Physical Activity Epidemiology

In recent years, a rapidly growing body of promising approaches has been achieved in physical activity research including approaches for disease prevention and the development of more advanced devices for physical activity assessment. Physical activity guidelines and intervention programs for the promotion of physical activity are based on the most updated research knowledge in physical activity epidemiology. There are still several unresolved questions that need to be addressed in future research to enhance our knowledge in this field and to incorporate that knowledge into effective physical activity intervention programs. To broaden our understanding of physical activity epidemiology, enhanced research on physical fitness and sedentary behavior is desirable, the latter of which is only beginning to accumulate available data. Moreover, we need to expand research on already existing approaches, such as the health-enhancing benefits of physical activity engagement, but we must also follow new promising theories and concepts. There is a strong research imperative of developing more sophisticated devices for physical activity assessment that allow moving away from subjective to objective physical activity measurements. Moreover, to establish appropriate and effective intervention programs, more emphasis should be placed on the exploration of determinants of physical activity, biological mechanisms linking physical activity to disease etiology, and physical activity behavior in population subgroups. This chapter summarizes areas of potential future research in physical activity epidemiology.

49.10.1 Improvements in Physical Activity Questionnaires

Despite their limited validity, PAQs are crucial in physical activity epidemiology as they represent an inexpensive and quickly administered tool to assess physical activity in large sample sizes at a relatively low participant burden. Compared to objective physical activity assessment instruments such as accelerometers, PAQs offer an advantage by providing a comprehensive measure of physical activity according to its types and components. Specifically, PAQs enable assessments of incline walking, water-based activities, and upper-body movements, such as rowing, activities that are not fully captured by uniaxial and triaxial accelerometers. Moreover, when comparing PAQs and accelerometers, PAQs better distinguish between closely related types of physical activities, such as light dancing and walking. Although considerable advances have been made in the area of questionnaire-based physical activity assessment during the past decades, future improvements in PAQs are needed (Matthews et al. 2012). Given the quantitatively significant contribution of light and moderate-intensity activities to overall energy expenditure and the potential gains in health provided by such activities, the methods used for precisely assessing the individual components of light and moderate activities should be given particular attention. For example, the cumulative effect of moderate-intensity activities, such as walking, may encompass a greater energetic impact on energy expenditure than the cumulative effect of short bursts of high-intensity sports activities (Westerterp 2001). In addition, more information is needed regarding which activities or combinations of activities performed in free-living individuals best meet the current physical activity recommendations. For example, it is currently not known whether leisure-time purposeful walking for 30 min a day on 5 days of the week better characterizes the aspect of physical activity that is relevant for disease prevention than the cumulative weekly amount of casual walking and stair climbing performed during housework, while shopping, and at the workplace.

49.10.2 Increased Use of Objective Physical Activity Assessments

Only a small number of available epidemiological investigations have used objective measures of physical activity, such as pedometers or accelerometers, to assess the association between physical activity and disease. However, objective devices show vast potential to provide more valid and reproducible assessments of physical activity levels than questionnaire-based measurements (Warren et al. 2010). They bear the limitation that activity from a previous time in life cannot be measured. As equipment costs continue to fall, these devices could reach widespread application in epidemiological research and population surveillance. New technologies, such as mobile phones, smart phones, interactive voice response, and interactive video games, also hold promise to become valuable tools for public health promotion of physical activity and delivery of physical activity interventions (van den Berg et al. 2007).

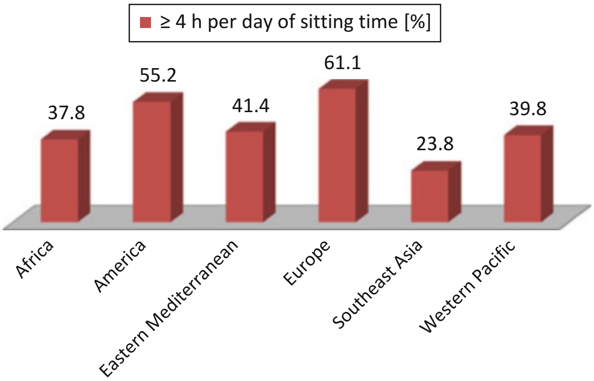
49.10.3 Measurement of Physical Fitness

Little is known about the impact of physical fitness on disease prevention. The sparse data available suggest that fitness is inversely related to risk of heart disease, cancer, and mortality (Blair et al. 1989). Some evidence indicates that physical fitness may be a more important determinant of health outcomes than physical activity, although this may be due to the imprecision in measuring physical activity as compared with the ability of assessing physical fitness. For example, in a study among men with type 2 diabetes, physically inactive men had a 1.7-fold (95% CI = 1.2–2.3) greater risk of all-cause mortality than physically active men, but the low-fitness group had a 2.1-fold (95% CI = 1.5–2.9) higher risk of all-cause mortality compared to fit men (Wei et al. 2000). In another study, a comparison of physical activity and cardiorespiratory fitness in relation to all-cause mortality revealed a stronger association of cardiorespiratory fitness than physical activity with all-cause mortality (Lee et al. 2011). Physical fitness represents an objective and reproducible measure that reflects recent physical activity habits and is not subject to recall bias and is also less likely prone to misclassification, behavior reactivity, and daily variation than physical activity. Thus, physical fitness may represent a superior measure or may provide additional information to studies on the relationships of physical activity to disease outcomes.

49.10.4 Assessment of Sedentary Behavior

From a societal and individual health-care perspective, being sedentary is currently perceived as the norm. Recently, it has been shown that 41.5% of adults worldwide spend 4 or more hours per day in sitting pursuits. As shown in Fig. 49.21, the time spent in sedentary behaviors varies considerably between WHO regions (Hallal et al. 2012). Sedentary behavior enhances the risk of heart disease, stroke, diabetes,

Fig. 49.21 Proportion (%) of individuals aged ≥15 years engaging in sedentary behavior by World Health Organization region (Hallal et al. 2012)



and cancer by 20% to 30% and decreases life span by up to 5 years (Wen et al. 2011). Thus, sedentary behavior places an enormous public health burden on society through burgeoning health-care costs and loss of productivity. Hu et al. (2003) examined the association between various sedentary behaviors and the risk of obesity and type 2 diabetes in women (Hu et al. 2003). For each 2 hour per day increase in time spent watching television, the risk of obesity was increased by 23% (95% CI = 17%–30%), and the risk of diabetes was increased by 14% (95% CI = 5%–23%). Sitting at work or standing was also associated with an increased risk of obesity and diabetes, albeit to a lesser extent than television watching (Hu et al. 2003). Gierach et al. (2009) observed an increased risk of endometrial cancer in women who sat 7 and more hours per day versus those who sat less than 3 h per day (*OR* = 1.56; 95% CI = 1.22–1.99).

Katzmarzyk and Lee (2012) evaluated the impact of sitting and television viewing on life expectancy in the U.S. population. Their analyses revealed that life expectancy in the U.S. population would be prolonged by 2 years if adults reduced their sitting time to less than 3 h per day and by 1.38 years if they reduced their television watching to less than 2 h per day. Thus, public health research and practical efforts to decrease sedentary behavior and to better understand the potentially adverse effects of sedentary behavior on disease risk are a priority (Thorp et al. 2011).

Sedentary behavior and physical activity have to be considered as two distinct concepts that may affect health independently. For instance, in high-income countries, persons are becoming more sedentary during their working time but are increasing their physical activity participation during recreation time. More research investigating the effect of this increasingly widespread physical activity pattern on health outcomes is needed.

49.10.5 Assessment of Physical Activity over the Life Course

An important area of future development in physical activity epidemiology is born from the need for increased knowledge of physical activity over the life course in relation to the maintenance of health and the prevention of disease.

It is unclear whether the potential for chronic disease risk reduction associated with increased physical activity is independent of the timing of physical activity or whether physical activity that is performed during potentially susceptible life stages is the relevant time period of potential risk reduction. For example, one study of timing of physical activity in relation to postmenopausal breast cancer risk among women aged 50 to 70 years of age found that increased levels of physical activity performed during the past 10 years were associated with decreased risk of postmenopausal breast cancer. By comparison, physical activity performed during the ages of 15 to 18, 19 to 29, and 35 to 39 years showed no relation to postmenopausal breast cancer reduction (Peters et al. 2009). Those data support the hypothesis that individual phases of hormonal change have distinct biological consequences that may differentially impact upon the association between physical activity and some types of cancers. There may also be distinct effects of physical activity levels at various stages during normal cell growth to diagnosed cancer. For example, Giovannucci et al. (1998) found no association between physical activity and total prostate cancer but reported a significant reduction in risk for metastatic prostate cancer comparing the highest with the lowest physical activity levels.

Similarly, physical activity relations may apply to further disease entities, such as heart disease and diabetes. For example, Wannamethee et al. (1998) investigated changes in physical activity in relation to the incidence of heart disease in older men. The men reported their physical activity participation once in 1978–1980 (Q1) and again 12 to 14 years later in 1992 (Q92). Those who were sedentary at the first point in time and who began participation in at least light activity by Q92 had a significant lower all-cause mortality risk than those who remained sedentary (*RR* = 0.55; 95% CI = 0.36–0.84). Moreover, change to a more active behavior improved both cardiovascular mortality (*RR* = 0.66; 95% CI = 0.35–1.23) and non-cardiovascular mortality (*RR* = 0.48; 95% CI = 0.27–0.85) (Wannamethee et al. 1998).

Thus, to tease out time-specific effects, more lifetime physical activity studies and investigations of changes of physical activity from sedentary behavior to more active behavior across different levels of physical activity are needed. Ideally, a comprehensive assessment of physical activity in a cohort study includes the measurement of levels of physical activity at various points in time. Moreover, RCTs and experimental studies are necessary to investigate the mechanisms by which physical activity timing influences disease risk.

49.10.6 Physical Activity in Tertiary Prevention of Disease

For some diseases such as cancer, the majority of physical activity research has been conducted on the primary prevention of disease, and less epidemiological evidence is available on the role of physical activity in tertiary prevention. For example, only a few studies have addressed the association between physical activity and mortality among survivors of cancers, such as prostate cancer, ovarian cancer, and colorectal cancer (Ballard-Barbash et al. 2012). Survivors of cancer or cardiovascular events have a higher risk of recurrence of disease and mortality (Chen et al. 2011; Roger et al. 2012). Moreover, it was reported that exercise may improve quality of life, cancer-related fatigue, and physical functioning in cancer survivors (Schmitz et al. 2010).

Survivors of cancer or a cardiac event may be more likely motivated to change their lifestyle along with enhanced physical activity participation than the general population. Thus, more physical activity research in survivors of chronic diseases is needed to guide physical activity recommendations on the secondary and tertiary prevention of diseases.

49.10.7 Assessment of Determinants of Physical Activity

The determinants and correlates of physical activity are complex, and they involve numerous factors, including age, sex, income, geographical location, transportation environment, and social and psychological factors such as self-efficacy and aesthetics. Identifying evidence-based personal and environmental determinants of physical activity represents an important step in designing appropriate intervention programs leading to changes in physical activity behavior (Trost et al. 2002). Future research requires enhanced assessments of determinants of physical activity in both high-income and low-income groups that incorporate cultural and country-level factors. Specifically, an increased knowledge of psychosocial, environmental, and economic determinants of transportation and recreational activity in low and middle income groups is needed to facilitate contextually tailored intervention programs aimed at thwarting the rapid development of a sedentary lifestyle brought about by increased urbanization, motorized commuting, and passive entertainment. Studying potential differences in physical activity determinants between low- and high-income groups might also help identify appropriate targets for intervention. For example, socioeconomic status may be inversely associated with total physical activity in low-income countries where transportation and occupational activity are the predominant domains of activity. By comparison, total physical activity may be positively associated with social class in high-income countries where recreational activity predominates.

49.10.8 Study of Physical Activity Levels in Population Subgroups

Considerable disparities exist in levels of physical activity in population subgroups. For example, Levine (2011) reported the greatest sedentariness being in the poorest countries. The author suggested several reasons for this observation, including violence that prevents from engaging in outdoor activities in poor regions, less availability of parks and sports facilities to people living in poor counties, and limited affordability of gym memberships and/or exercise equipment for persons who live in poverty. In future research, more attention should be paid to subgroups of individuals living in countries of low and middle income, socially disadvantaged individuals, ethnic/racial minorities, and persons with disabilities. Moreover, in light of the obesity epidemic, more attention should be paid to obese persons who are at risk for low levels of physical activity. Emphasis should be placed on research to clarify from which physical activity interventions these population subgroups will benefit most. This represents an important step in devising personalized disease prevention strategies. Individuals of these subgroups are more frequently exposed to social and/or environmental barriers to physical activity and may have fewer possibilities to overcome such obstacles than other population groups. Strategies need to be developed to reach these subgroups and to facilitate their access to intervention programs, for example, by mobilizing social networks (Marcus et al. 2006). When tailoring physical activity programs for weight loss and weight control in overweight and obese persons, we should bear in mind that being overweight and having sedentary habits is a barrier to becoming physically active (Sherwood and Jeffery 2000). To achieve weight loss, engagement in a higher amount of physical activity is required which might be difficult for such persons to perform. Moreover, overweight or obese individuals may additionally be burdened by physiological, behavioral, and psychological risk factors (Sherwood and Jeffery 2000). In addition, creation of a supportive environment to facilitate physical activity participation, including neighborhoods, parks, and recreation facilities, is imperative to create effective interventions for the promotion of physical activity. Thus, physical activity epidemiologists and public health professionals need to develop physical activity programs and policy initiatives to enhance physical activity opportunities in the entire population, with particular attention paid to efforts aimed at meeting the needs of population subgroups with limited access to physical activity.

49.10.9 Creation of Strategic Partnerships to Influence Physical Environments

Although an enhanced placement of physical activity within the area of public health appears to be important, a larger potential to increase levels of physical activity in the population might be changing the physical environment (Brownson et al. 2001). This is best accomplished through the creation of strategic alliances with sectors beyond public health. Thus, partnerships between policy makers and public health, behavioral, economic, and social scientists, specifically in the areas of transport and urban planning, should be developed. A systems-based approach incorporating qualitative and quantitative physical activity research with policy analysis is required to evaluate the multifaceted conditions that determine population levels of physical activity. Such data are needed to create the basis for building physical environments in which the opportunity to engage in a physically active lifestyle is easy, convenient, affordable, and safe. Cross-sectoral efforts that promote individual and population-level physical activity as a co-benefit are already in place, such as the creation of green spaces, pedestrian and bicycle systems, integrated transportation networks, traffic restriction, and taxation of petrol.

49.10.10 Examination of Biological Mechanisms Linking Physical Activity to Health

Given evidence of a protective effect of physical activity on disease risk and the potential existence of numerous biological mechanisms linking increased physical activity to decreased risk of disease, little is known about the exact biological pathways through which physical activity influences biological processes leading to protection from disease and improvements in prognosis (McTiernan 2008). Yet a thorough understanding of the biological mechanisms is pivotal to developing specific molecular targets for disease prevention and survival. In addition, the relation of the different components of physical activity to biological pathways remains to be defined. For example, it is not known whether a reduction in low-grade systemic chronic inflammation is best achieved by low-intensity activities such as walking or if vigorous activities such as jogging are required. The sparse data available suggest the existence of a "J"-shaped relationship between physical activity and chronic inflammation based on the commonly observed increase in the frequency of upper respiratory tract infections among elite endurance athletes during periods of intensive training (Walsh et al. 2011).

49.10.11 Expansion of Lifestyle Physical Activity Intervention Programs

Based on research studies and public health guidelines (Epstein et al. 1990; Blair 1993; U.S. Department of Health and Human Services 1996), Dunn et al. (1998) coined the term lifestyle physical activity as "…the daily accumulation of at least 30 minutes of self-selected activities, which includes all leisure, occupational, or household activities that are at least moderate to vigorous in their intensity and could be planned or unplanned activities that are part of everyday life." Thus, a lifestyle physical activity approach does not solely target persons who are interested in doing exercise training and sports, but also facilitates entry for a wider population, including sedentary individuals, to become more physically active thereby providing the potential for increased physical activity to improve health (Haskell 1994). Moreover, as time and access are the most frequently reported barriers to physical activity, the lifestyle activity approach may overcome these obstacles.

Most lifestyle activity interventions focus on approaches to behavior change in individuals and populations. Interventions are typically delivered by face-to-face contacts, telephone and mail, computer-mediated or web-based programs, or a combination of these components (Artinian et al. 2010). Further approaches include modifications of the environment, such as putting up signs to increase stair climbing (Dunn et al. 1998).

Although effective strategies regarding lifestyle activity interventions have been reported in the past (Dunn et al. 1998), there are still issues regarding the promotion of lifestyle activity interventions that need to be targeted in future research. Those include the development of lifestyle intervention programs in different population groups to determine which intervention is best effective for which group. Furthermore, new technologies such as text messaging and social networking for promotion of lifestyle activity should be taken into account (Artinian et al. 2010). Follow-ups of intervention programs are necessary to explore their efficacy with regards to long-term lifestyle activity maintenance (Dunn et al. 1998). Additionally, both behavioral interventions and environmental manipulations need to be tested for their cost-effectiveness. Lifestyle physical activity intervention strategies that are effective under real-life conditions are of major importance, and an understanding of how to embed these strategies in health practice along with organizational and policy issues is an important concern that requires consideration (Artinian et al. 2010).

49.10.12 Evaluation of Vigorous Activity to Promote Body Leanness

While the 1995 ACSM/CDC physical activity recommendation posits the promotion of moderate physical activity (Pate et al. 1995), the 2007 and 2011 ACSM guidelines additionally include recommendations on vigorous-intensity activities for muscle strength and bone health (Haskell et al. 2007; Garber et al. 2011). It has been shown that vigorous physical activity exerts beneficial effects on overall and visceral fat, and bone mass (Kohrt et al. 2004; Friedenreich et al. 2011). Recently, a new theory has emerged which hypothesizes that vigorous physical activity promotes the differentiation of immature stem cells into lean tissue cells rather than into fat cells through mechanical stimulation (Luu et al. 2009). In mice, it was found that low-magnitude mechanical signals promote bone development while inhibiting the development of fat mass. Moreover, in 15- to 20-year-old women with osteopenia, similar observations have been made with mechanical stimulation, which led to an increased bone mass and reduced visceral fat in the lumbar area (Luu et al. 2009). Further research on vigorous activity is needed in support of this theory, which may represent a basis to simultaneously prevent obesity and osteoporosis.

49.11 Conclusions

Physical activity epidemiology represents a young scientific discipline which has emerged over the past decades. It represents a research area concerned with understanding the associations between physical activity and morbidity and mortality, and its interrelations with other health factors. The goal of physical activity epidemiology is to apply that knowledge to intervention programs and to inform recommendations for the prevention and control of diseases and the promotion of health.

There is mounting evidence for benefits of physical activity on mortality, CVD, various types of cancer, type 2 diabetes, and other diseases. However, the recent descriptive epidemiology of physical activity has revealed a decline in overall physical activity participation along with a steady increase in sedentary behavior among populations of industrialized countries. Physical activity research describes a dynamic process continuously building new knowledge about the effect of physical activity on health of which the best evidence provides the basis for physical activity recommendations and intervention programs. For the past two decades, there has been consensus among several governmental bodies and organizations on the recommendation of engaging in at least 30 minutes of moderate physical activity on most days of the week for beneficial health. Engagement in vigorous-intensity activity and resistance training for major muscle groups several times during the week are also elements of the current recommendations. The lifestyle physical activity approach describes a promising strategy to overcome time and access barriers to physical activity and to reach a wider population including sedentary individuals.

Strong research designs should provide the optimal basis for guidelines and effective physical activity intervention programs. Physical activity is a complex and multifaceted human behavior with large day-to-day variability, and its accurate quantification remains a challenge in physical activity research. Before conducting analyses of physical activity, a number of issues need to be considered including considerations regarding methodological aspects (e.g., study design, confounding, reverse causation), the nature of physical activity assessment (e.g., subjective or objective measures), dimensions of physical activity (e.g., intensity, duration, frequency), and population characteristics (e.g., age, sex, geographical location). The selection of a specific study design depends on the research question, and it represents a balance between the needs for accuracy and feasibility. Observational epidemiological studies allow an assessment of physical activity patterns under free-living conditions on a large scale. Information on potential biological mechanisms linking physical activity to diverse disease outcomes is best derived from RCTs and laboratory experiments.

Physical activities of light to moderate intensity, which are often performed as part of daily routine activities and reflect the nature of physical activity among children and the elderly, are more difficult to measure and cannot easily be captured by standard physical activity measurement methods. Physical activity questionnaires allow the assessment of movement in free-living conditions in large samples, albeit at the expense of accuracy. Objectively measured physical activity by accelerometers provides more accurate measurements yet can be limited by cost implications in large-scale investigations and does not provide retrospective physical activity information. DLW and indirect calorimetry represent the most accurate instruments for assessment of energy expended, but they are limited to controlled laboratory settings and to small numbers of individuals and thus solely serve as criterion methods.

Methodological considerations, including appropriate control for confounding, and causality issues are important in epidemiological studies to infer meaningful associations between physical activity levels and health outcomes. Repeated measurements of physical activity during follow-up reduce measurement errors and allow detecting sensitive time frames in disease development in relation to physical activity exposure.

Although considerable progress in physical activity research has been made in recent years, many unresolved questions in physical activity epidemiology require further attention. Besides the development of more sophisticated devices of physical activity measurement, moving from laboratory settings to large-scale studies of physical activity on disease outcomes across all ages under free-living conditions is needed. Furthermore, analyses of those at risk for low engagement in physical activity, such as people with disabilities, obese individuals, ethnic/racial minorities, or socially disadvantaged persons, are encouraged to establish personalized disease prevention strategies. Little is known about biomarkers and etiological pathways through which physical activity affects health outcomes. Thus, improving our understanding of determinants of physical activity and identifying biological mechanisms may help devise intervention strategies for the prevention of diseases and premature mortality. Moreover, enhanced research on physical fitness is required, which may provide additional information about the health effects of physical activity. Our public health goal is moving from sedentariness to a more active population. Thus, in addition to physical activity pursuits, more attention should be paid to time spent in sedentary behaviors in relation to health effects. Finally, while continuing the efforts to build the evidence basis on health effects of physical activity and sedentary behaviors, building a bridge between physical activity research and policy makers and scientists of related research fields is warranted to facilitate behavior modification and environmental change for the promotion of increased physical activity.

References

About 10,000 Steps Project (2012) 10,000 Steps Rockhampton Project, Queensland, Australia. http://www.10000steps.org.au. Accessed 1 Nov 2012

Acheson KJ, Campbell IT, Edholm OG, Miller DS, Stock MJ (1980) The measurement of daily energy expenditure – an evaluation of some techniques. Am J Clin Nutr 33:1155–1164

ACSM/CDC (1995) Pate RR, Pratt M, Blair SN et al. Physical activity and public health. A recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine. JAMA. 273:402–407

Adami PE, Negro A, Lala N, Martelletti P (2010) The role of physical activity in the prevention and treatment of chronic diseases. Clin Ter 161:537–541

Adams J (2006) Trends in physical activity and inactivity amongst US 14–18 year olds by gender, school grade and race, 1993–2003: evidence from the youth risk behavior survey. BMC Public Health 6:57. doi:10.1186/1471-2458-6-57

Ainsworth BE, Haskell WL, Herrmann SD, Meckes N, Bassett DR Jr, Tudor-Locke C, Greer JL, Vezina J, Whitt-Glover MC, Leon AS (2011) Compendium of physical activities: a second update of codes and MET values. Med Sci Sports Exerc 43:1575–1581

Armstrong N, Welsman JR (2006) The physical activity patterns of European youth with reference to methods of assessment. Sports Med 36:1067–1086

Artinian NT, Fletcher GF, Mozaffarian D, Kris-Etherton P, Van Horn L, Lichtenstein AH, Kumanyika S, Kraus WE, Fleg JL, Redeker NS, Meininger JC, Banks J, Stuart-Shor EM, Fletcher BJ, Miller TD, Hughes S, Braun LT, Kopin LA, Berra K, Hayman LL, Ewing LJ, Ades PA, Durstine JL, Houston-Miller N, Burke LE (2010) Interventions to promote physical activity and dietary lifestyle changes for cardiovascular risk factor reduction in adults: a scientific statement from the American Heart Association. Circulation 122:406–441

Australian Government Department of Health and Ageing (2010) Physical activity guidelines. http://www.health.gov.au/internet/main/publishing.nsf/Content/health-pubhlth-strateg-phys-act-guidelines#reasons. Accessed 8 Aug 2012

Baecke JA, Burema J, Frijters JE (1982) A short questionnaire for the measurement of habitual physical activity in epidemiological studies. Am J Clin Nutr 36:936–942

Ballard-Barbash R, Friedenreich CM, Courneya KS, Siddiqi SM, McTiernan A, Alfano CM (2012) Physical activity, biomarkers, and disease outcomes in cancer survivors: a systematic review. J Natl Cancer Inst 104:815–840

Baranowski T (1985) Methodologic issues in self-report of health behavior. J Sch Health 55:179–182

Barreira TV, Kang M, Caputo JL, Farley RS, Renfrow MS (2009) Validation of the Actiheart monitor for the measurement of physical activity. Int J Exerc Sci 2:60–71

Bassett DR Jr, Ainsworth BE, Leggett SR, Mathien CA, Main JA, Hunter DC, Duncan GE (1996) Accuracy of five electronic pedometers for measuring distance walked. Med Sci Sports Exerc 28:1071–1077

Bassett DR Jr, Ainsworth BE, Swartz AM, Strath SJ, O'Brien WL, King GA (2000) Validity of four motion sensors in measuring moderate intensity physical activity. Med Sci Sports Exerc 32:S471–S480

Bassett DR Jr, Rowlands A, Trost SG (2012) Calibration and validation of wearable monitors. Med Sci Sports Exerc 44:S32–S38

Bauman AE, Sallis JF, Dzewaltowski DA, Owen N (2002) Toward a better understanding of the influences on physical activity: the role of determinants, correlates, causal variables, mediators, moderators, and confounders. Am J Prev Med 23:5–14

Bauman A, Ainsworth BE, Sallis JF, Hagstromer M, Craig CL, Bull FC, Pratt M, Venugopal K, Chau J, Sjostrom M (2011) The descriptive epidemiology of sitting. A 20-country comparison using the International Physical Activity Questionnaire (IPAQ). Am J Prev Med 41:228–235

Berlin JE, Storti KL, Brach JS (2006) Using activity monitors to measure physical activity in free-living conditions. Phys Ther 86:1137–1145

Besson H, Harwood CA, Ekelund U, Finucane FM, McDermott CJ, Shaw PJ, Wareham NJ (2010) Validation of the historical adulthood physical activity questionnaire (HAPAQ) against objective measurements of physical activity. Int J Behav Nutr Phys Act 7:54

Blair SN (1993) C.H. McCloy Research Lecture: physical activity, physical fitness, and health. Res Q Exerc Sport 64:365–376

Blair SN, Kohl HW 3rd, Paffenbarger RS Jr, Clark DG, Cooper KH, Gibbons LW (1989) Physical fitness and all-cause mortality. A prospective study of healthy men and women. JAMA 262:2395–2401

Blair SN, Dowda M, Pate RR, Kronenfeld J, Howe HG Jr, Parker G, Blair A, Fridinger F (1991) Reliability of long-term recall of participation in physical activity by middle-aged men and women. Am J Epidemiol 133:266–275

Blair SN, Cheng Y, Holder JS (2001) Is physical activity or physical fitness more important in defining health benefits? Med Sci Sports Exerc 33:S379–S399; discussion S419–S320

Boon RM, Hamlin MJ, Steel GD, Ross JJ (2010) Validation of the New Zealand Physical Activity Questionnaire (NZPAQ-LF) and the International Physical Activity Questionnaire (IPAQ-LF) with accelerometry. Br J Sports Med 44:741–746

Booth ML, Okely AD, Chey T, Bauman A (2002) The reliability and validity of the adolescent physical activity recall questionnaire. Med Sci Sports Exerc 34:1986–1995

Borg G (1998) Borg's perceived exertion and pain scales. Human Kinetics, Campaign

Borodulin K, Laatikainen T, Juolevi A, Jousilahti P (2008) Thirty-year trends of physical activity in relation to age, calendar time and birth cohort in Finnish adults. Eur J Public Health 18:339–344

Bouchard C, Tremblay A, Leblanc C, Lortie G, Savard R, Theriault G (1983) A method to assess energy expenditure in children and adults. Am J Clin Nutr 37:461–467

Bouchard C, Daw EW, Rice T, Perusse L, Gagnon J, Province MA, Leon AS, Rao DC, Skinner JS, Wilmore JH (1998) Familial resemblance for VO2max in the sedentary state: the HERITAGE Family Study. Med Sci Sports Exerc 30:252–258

Bouchard C, An P, Rice T, Skinner JS, Wilmore JH, Gagnon J, Perusse L, Leon AS, Rao DC (1999) Familial aggregation of VO(2max) response to exercise training: results from the HERITAGE Family Study. J Appl Physiol 87:1003–1008

Bowles HR, Fitzgerald SJ, Morrow JR Jr, Jackson AW, Blair SN (2004) Construct validity of self-reported historical physical activity. Am J Epidemiol 160:279–286

Brage S, Wedderkopp N, Franks PW, Andersen LB, Froberg K (2003) Reexamination of validity and reliability of the CSA monitor in walking and running. Med Sci Sports Exerc 35: 1447–1454

Brage S, Brage N, Franks PW, Ekelund U, Wong MY, Andersen LB, Froberg K, Wareham NJ (2004) Branched equation modeling of simultaneous accelerometry and heart rate monitoring improves estimate of directly measured physical activity energy expenditure. J Appl Physiol 96:343–351

Brage S, Brage N, Franks PW, Ekelund U, Wareham NJ (2005) Reliability and validity of the combined heart rate and movement sensor Actiheart. Eur J Clin Nutr 59:561–570

Brown WH, Pfeiffer KA, McLver KL, Dowda M, Almeida MJ, Pate RR (2006) Assessing preschool children's physical activity: the Observational System for Recording Physical Activity in children-preschool version. Res Q Exerc Sport 77:167–176

Brownson RC, Baker EA, Housemann RA, Brennan LK, Bacak SJ (2001) Environmental and policy determinants of physical activity in the United States. Am J Public Health 91: 1995–2003

Brownson RC, Boehmer TK (2004) Patterns and trends in physical activity, occupation, transportation, land use, and sedentary behaviors. TRB Special Report 282. Does the built environment influence physical activity? Examining the evidence (Paper prepared for the Transportation Research Board and the Institute of Medicine Committee on Physical activity, Health, Transportation, and Land Use. http://trb.org/downloads/sr282papers/sr282Brownson.pdf. Accessed 2 Aug 2012

Brownson RC, Boehmer TK, Luke DA (2005) Declining rates of physical activity in the United States: what are the contributors? Annu Rev Public Health 26:421–443

Bucksch J (2005) Physical activity of moderate intensity in leisure time and the risk of all cause mortality. Br J Sports Med 39:632–638

Bucksch J, Schlicht W (2006) Health-enhancing physical activity and the prevention of chronic diseases – An epidemiological review. Sozial- und Präventivmedizin 51:281–301

Canadian Society for Exercise Physiology (2012) Canadian physical activity guidelines and Canadian sedentary behaviour guidelines. http://www.csep.ca/english/view.asp?x=804. Accessed 8 Aug 2012

Caspersen CJ (1989) Physical activity epidemiology: concepts, methods, and applications to exercise science. Exerc Sport Sci Rev 17:423–473

Caspersen CJ, Powell KE, Christenson GM (1985) Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. Public Health Rep 100:126–131

Centers for Disease Control and Prevention (2009) Behavioral risk factor surveillance system. Prevalence and trends data. http://apps.nccd.cdc.gov/BRFSS/page.asp?cat=XX&yr=2009&state=All#XX. Accessed 10 Aug 2012

Centers for Disease Control and Prevention (2011) Youth Risk Behavior Surveillance System (YRBSS). http://www.cdc.gov/healthyyouth/yrbs/pdf/us_overview_yrbs.pdf. Accessed 10 Aug 2012

Centers for Disease Control and Prevention (2012) Surveillance. http://www.cdc.gov/physicalactivity/data/surveillance.html. Accessed 10 Aug 2012

Chen KY, Bassett DR Jr (2005) The technology of accelerometry-based activity monitors: current and future. Med Sci Sports Exerc 37:S490–S500

Chen MH, Colan SD, Diller L (2011) Cardiovascular disease: cause of morbidity and mortality in adult survivors of childhood cancers. Circ Res 108:619–628

Chodzko-Zajko WJ, Proctor DN, Fiatarone Singh MA, Minson CT, Nigg CR, Salem GJ, Skinner JS (2009) American College of Sports Medicine position stand. Exercise and physical activity for older adults. Med Sci Sports Exerc 41:1510–1530

Chomistek AK, Cook NR, Flint AJ, Rimm EB (2012) Vigorous-intensity leisure-time physical activity and risk of major chronic disease in men. Med Sci Sports Exerc 44:1898–1905

Church TS, Thomas DM, Tudor-Locke C, Katzmarzyk PT, Earnest CP, Rodarte RQ, Martin CK, Blair SN, Bouchard C (2011) Trends over 5 decades in U.S. occupation-related physical activity and their associations with obesity. PLoS One 6:e19657

Colditz GA, Cannuscio CC, Frazier AL (1997) Physical activity and reduced risk of colon cancer: implications for prevention. Cancer Causes Control 8:649–667

Courneya KS, Friedenreich CM (2011) Physical activity and cancer. Springer, Heidelberg

Courneya KS, Keats MR, Turner AR (2000) Physical exercise and quality of life in cancer patients following high dose chemotherapy and autologous bone marrow transplantation. Psychooncology 9:127–136

Craig CL, Marshall AL, Sjostrom M, Bauman AE, Booth ML, Ainsworth BE, Pratt M, Ekelund U, Yngve A, Sallis JF, Oja P (2003) International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 35:1381–1395

Crouter SE, Schneider PL, Karabulut M, Bassett DR Jr (2003) Validity of 10 electronic pedometers for measuring steps, distance, and energy cost. Med Sci Sports Exerc 35:1455–1460

Crouter SE, Churilla JR, Bassett DR Jr (2006a) Estimating energy expenditure using accelerometers. Eur J Appl Physiol 98:601–612

Crouter SE, Clowers KG, Bassett DR Jr (2006b) A novel method for using accelerometer data to predict energy expenditure. J Appl Physiol 100:1324–1331

Department of Health, Education and Welfare (1964) Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service Washington DC: U.S.

Dishman RK, Washburn RA, Heath GW (2004) Physical activity epidemiology. Human Kinetics, Inc, Champaign

Dollman J, Okely AD, Hardy L, Timperio A, Salmon J, Hills AP (2009) A hitchhiker's guide to assessing young people's physical activity: deciding what method to use. J Sci Med Sport 12:518–525

Dunn AL, Andersen RE, Jakicic JM (1998) Lifestyle physical activity interventions. History, short- and long-term effects, and recommendations. Am J Prev Med 15:398–412

Epstein LH, Valoski A, Wing RR, McCurley J (1990) Ten-year follow-up of behavioral, family-based treatment for obese children. JAMA 264:2519–2523

European Commission (2008) EU physical activity guidelines. Recommended policy actions in support of health-enhancing physical activity. http://ec.europa.eu/sport/library/doc/c1/pa_guidelines_4th_consolidated_draft_en.pdf. Accessed 8 Aug 2012

European Commission (2012) EU Expert Groups 'Sport, Health and Participation' and 'Sustainable Financing of Sport': reports and first deliverables available. http://ec.europa.eu/sport/news/20120803-eu-xg-shp-fin-rpt_en.htm. Accessed 8 Aug 2012

Evenson KR, Catellier DJ, Gill K, Ondrak KS, McMurray RG (2008) Calibration of two objective measures of physical activity for children. J Sports Sci 26:1557–1565

Fletcher GF, Balady GJ, Amsterdam EA, Chaitman B, Eckel R, Fleg J, Froelicher VF, Leon AS, Pina IL, Rodney R, Simons-Morton DA, Williams MA, Bazzarre T (2001) Exercise standards for testing and training: a statement for healthcare professionals from the American Heart Association. Circulation 104:1694–1740

Friberg E, Mantzoros CS, Wolk A (2006) Physical activity and risk of endometrial cancer: a population-based prospective cohort study. Cancer Epidemiol Biomark Prev 15:2136–2140

Friedenreich CM (2001) Physical activity and cancer prevention: from observational to intervention research. Cancer Epidemiol Biomark Prev 10:287–301

Friedenreich CM, Cust AE (2008) Physical activity and breast cancer risk: impact of timing, type and dose of activity and population subgroup effects. Br J Sports Med 42:636–647

Friedenreich CM, Bryant HE, Courneya KS (2001) Case-control study of lifetime physical activity and breast cancer risk. Am J Epidemiol 154:336–347

Friedenreich C, Cust A, Lahmann PH, Steindorf K, Boutron-Ruault MC, Clavel-Chapelon F, Mesrine S, Linseisen J, Rohrmann S, Pischon T, Schulz M, Tjonneland A, Johnsen NF, Overvad K, Mendez M, Arguelles MV, Garcia CM, Larranaga N, Chirlaque MD, Ardanaz E, Bingham S, Khaw KT, Allen N, Key T, Trichopoulou A, Dilis V, Trichopoulos D, Pala V, Palli D, Tumino R, Panico S, Vineis P, Bueno-de-Mesquita HB, Peeters PH, Monninkhof E, Berglund G, Manjer J, Slimani N, Ferrari P, Kaaks R, Riboli E (2007) Physical activity and risk of endometrial cancer: the European prospective investigation into cancer and nutrition. Int J Cancer 121:347–355

Friedenreich CM, Woolcott CG, McTiernan A, Terry T, Brant R, Ballard-Barbash R, Irwin ML, Jones CA, Boyd NF, Yaffe MJ, Campbell KL, McNeely ML, Karvinen KH, Courneya KS (2011) Adiposity changes after a 1-year aerobic exercise intervention among postmenopausal women: a randomized controlled trial. Int J Obes (Lond) 35:427–435

Garber CE, Blissmer B, Deschenes MR, Franklin BA, Lamonte MJ, Lee IM, Nieman DC, Swain DP (2011) American College of Sports Medicine position stand. Quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults: guidance for prescribing exercise. Med Sci Sports Exerc 43:1334–1359

Gierach GL, Chang SC, Brinton LA, Lacey JV Jr, Hollenbeck AR, Schatzkin A, Leitzmann MF (2009) Physical activity, sedentary behavior, and endometrial cancer risk in the NIH-AARP Diet and Health Study. Int J Cancer 124:2139–2147

Giovannucci E, Leitzmann M, Spiegelman D, Rimm EB, Colditz GA, Stampfer MJ, Willett WC (1998) A prospective study of physical activity and prostate cancer in male health professionals. Cancer Res 58:5117–5122

Guthold R, Cowan MJ, Autenrieth CS, Kann L, Riley LM (2010) Physical activity and sedentary behavior among schoolchildren: a 34-country comparison. J Pediatr 157:43–49 e41

Hallal PC, Andersen LB, Bull FC, Guthold R, Haskell W, Ekelund U (2012) Global physical activity levels: surveillance progress, pitfalls, and prospects. Lancet 380:247–257

Harris TJ, Owen CG, Victor CR, Adams R, Ekelund U, Cook DG (2009) A comparison of questionnaire, accelerometer, and pedometer: measures in older people. Med Sci Sports Exerc 41:1392–1402

Haskell WL (1994) Health consequences of physical-activity - understanding and challenges regarding dose-response. Med Sci Sports Exerc 26:649–660

Haskell WL, Kiernan M (2000) Methodologic issues in measuring physical activity and physical fitness when evaluating the role of dietary supplements for physically active people. Am J Clin Nutr 72:541S–550S

Haskell WL, Lee IM, Pate RR, Powell KE, Blair SN, Franklin BA, Macera CA, Heath GW, Thompson PD, Bauman A (2007) Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. Circulation 116:1081–1093

Healy GN, Dunstan DW, Salmon J, Cerin E, Shaw JE, Zimmet PZ, Owen N (2007) Objectively measured light-intensity physical activity is independently associated with 2-h plasma glucose. Diabetes Care 30:1384–1389

Healy GN, Dunstan DW, Salmon J, Shaw JE, Zimmet PZ, Owen N (2008) Television time and continuous metabolic risk in physically active adults. Med Sci Sports Exerc 40:639–645

Heath GW, Parra DC, Sarmiento OL, Andersen LB, Owen N, Goenka S, Montes F, Brownson RC (2012) Evidence-based intervention in physical activity: lessons from around the world. Lancet 380:272–281

Hill AB (1965) The environment and disease: association or causation? Proc R Soc Med 58:295–300

Hoefler M (2005) The Bradford Hill considerations on causality: a counterfactual perspective. Emerg Themes Epidemiol 2. doi:10.1186/1742-7622-2-11

Howley ET (2001) Type of activity: resistance, aerobic and leisure versus occupational physical activity. Med Sci Sports Exerc 33:S364–S369; discussion S419–S320

Hu F (2008) Obesity epidemiology. Oxford University Press, New York

Hu FB, Li TY, Colditz GA, Willett WC, Manson JE (2003) Television watching and other sedentary behaviors in relation to risk of obesity and type 2 diabetes mellitus in women. JAMA 289: 1785–1791

Institute of Medicine of the National Academy (2002) Dietary reference intakes for energy, carbohydrate, fiber, fat, fatty acids, cholesterol, protein and amino acids (macronutrients). National Academy Press, Washington, DC

International Physical Activity Questionnaire (2012) Questionnaires. https://sites.google.com/site/theipaq/questionnaires. Accessed 25 Sept 2012

Janz KF (2002) Use of heart rate monitors to assess physical activity. In: Welk GJ (ed) Physical activity assessments for health-related research. Human Kinetics, Inc., Champaign, pp 143–161

Kaaks R, Lukanova A, Kurzer MS (2002) Obesity, endogenous hormones, and endometrial cancer risk: a synthetic review. Cancer Epidemiol Biomark Prev 11:1531–1543

Kahn EB, Ramsey LT, Brownson RC, Heath GW, Howze EH, Powell KE, Stone EJ, Rajab MW, Corso P (2002) The effectiveness of interventions to increase physical activity. A systematic review. Am J Prev Med 22:73–107

Katzmarzyk PT, Lee IM (2012) Sedentary behaviour and life expectancy in the USA: a cause-deleted life table analysis. BMJ Open 2:e000828. doi:10.1136/bmjopen-2012-000828

Kesaniemi YK, Danforth E Jr, Jensen MD, Kopelman PG, Lefebvre P, Reeder BA (2001) Dose-response issues concerning physical activity and health: an evidence-based symposium. Med Sci Sports Exerc 33:S351–S358

Kilanowski CK, Consalvi AR, Epstein LH (1999) Validation of an electronic pedometer for measurement of physical activity in children. Pediatr Exerc Sci 11:63–68

Knuth AG, Hallal PC (2009) Temporal trends in physical activity: a systematic review. J Phys Act Health 6:548–559

Kohrt WM, Bloomfield SA, Little KD, Nelson ME, Yingling VR (2004) American College of Sports Medicine Position Stand: physical activity and bone health. Med Sci Sports Exerc 36:1985–1996

Kolle E, Steene-Johannessen J, Andersen LB, Anderssen SA (2009) Seasonal variation in objectively assessed physical activity among children and adolescents in Norway: a cross-sectional study. Int J Behav Nutr Phys Act 6:36. doi:10.1186/1479-5868-6-36

Kwon S, Jamal M, Zamba GK, Stumbo P, Samuel I (2010) Validation of a novel physical activity assessment device in morbidly obese females. J Obes. doi:10.1155/2010/856376

Lamonte MJ, Ainsworth BE (2001) Quantifying energy expenditure and physical activity in the context of dose response. Med Sci Sports Exerc 33:S370–S378; discussion S419–S320

LaPorte RE, Montoye HJ, Caspersen CJ (1985) Assessment of physical activity in epidemiologic research: problems and prospects. Public Health Rep 100:131–146

Leatherdale ST, Wong SL (2008) Modifiable characteristics associated with sedentary behaviours among youth. Int J Pediatr Obes 3:93–101

Lee DC, Sui X, Ortega FB, Kim YS, Church TS, Winett RA, Ekelund U, Katzmarzyk PT, Blair SN (2011) Comparisons of leisure-time physical activity and cardiorespiratory fitness as predictors of all-cause mortality in men and women. Br J Sports Med 45:504–510

Lee IM, Blair NS, Manson J, Paffenbarger JS (2009) Epidemiological methods in physical activity studies. Oxford University Press, New York

Leitzmann MF, Rimm EB, Willett WC, Spiegelman D, Grodstein F, Stampfer MJ, Colditz GA, Giovannucci E (1999) Recreational physical activity and the risk of cholecystectomy in women. N Engl J Med 341:777–784

Leitzmann MF, Park Y, Blair A, Ballard-Barbash R, Mouw T, Hollenbeck AR, Schatzkin A (2007) Physical activity recommendations and decreased risk of mortality. Arch Intern Med 167:2453–2460

Leitzmann MF, Koebnick C, Abnet CC, Freedman ND, Park Y, Hollenbeck A, Ballard-Barbash R, Schatzkin A (2009) Prospective study of physical activity and lung cancer by histologic type in current, former, and never smokers. Am J Epidemiol 169:542–553

Levine JA (2011) Poverty and obesity in the U.S. Diabetes 60:2667–2668

Livingstone MB, Prentice AM, Coward WA, Ceesay SM, Strain JJ, McKenna PG, Nevin GB, Barker ME, Hickey RJ (1990) Simultaneous measurement of free-living energy expenditure by the doubly labeled water method and heart-rate monitoring. Am J Clin Nutr 52:59–65

Livingstone MB, Coward WA, Prentice AM, Davies PS, Strain JJ, McKenna PG, Mahoney CA, White JA, Stewart CM, Kerr MJ (1992) Daily energy expenditure in free-living children: comparison of heart-rate monitoring with the doubly labeled water (2H2(18)O) method. Am J Clin Nutr 56:343–352

Livingstone MB, Robson PJ, Wallace JM, McKinley MC (2003) How active are we? Levels of routine physical activity in children and adults. Proc Nutr Soc 62:681–701

Luu YK, Capilla E, Rosen CJ, Gilsanz V, Pessin JE, Judex S, Rubin CT (2009) Mechanical stimulation of mesenchymal stem cell proliferation and differentiation promotes osteogenesis while preventing dietary-induced obesity. J Bone Miner Res 24:50–61

MacAuley D (1994) A history of physical activity, health and medicine. J R Soc Med 87:32–35

Macfarlane DJ, Lee CC, Ho EY, Chan KL, Chan D (2006) Convergent validity of six methods to assess physical activity in daily life. J Appl Physiol 101:1328–1334

Mai PL, Sullivan-Halley J, Ursin G, Stram DO, Deapen D, Villaluna D, Horn-Ross PL, Clarke CA, Reynolds P, Ross RK, West DW, Anton-Culver H, Ziogas A, Bernstein L (2007) Physical activity and colon cancer risk among women in the California Teachers Study. Cancer Epidemiol Biomark Prev 16:517–525

Mann CJ (2003) Observational research methods. Research design II: cohort, cross sectional, and case-control studies. Emerg Med J 20:54–60

Manson JE, Hu FB, Rich-Edwards JW, Colditz GA, Stampfer MJ, Willett WC, Speizer FE, Hennekens CH (1999) A prospective study of walking as compared with vigorous exercise in the prevention of coronary heart disease in women. N Engl J Med 341:650–658

Marcus BH, Williams DM, Dubbert PM, Sallis JF, King AC, Yancey AK, Franklin BA, Buchner D, Daniels SR, Claytor RP (2006) Physical activity intervention studies: what we know and what we need to know: a scientific statement from the American Heart Association Council on Nutrition, Physical Activity, and Metabolism (Subcommittee on Physical Activity); Council on Cardiovascular Disease in the Young; and the Interdisciplinary Working Group on Quality of Care and Outcomes Research. Circulation 114:2739–2752

Matthews CE (2002) Use of self-report instruments to assess physical activity. In: Welk GJ (ed) Physical activity assessments for health-related research. Human Kinetics, Inc., Champaign, pp 107–123

Matthews CE, Hebert JR, Freedson PS, Stanek EJ 3rd, Merriam PA, Ebbeling CB, Ockene IS (2001a) Sources of variance in daily physical activity levels in the seasonal variation of blood cholesterol study. Am J Epidemiol 153:987–995

Matthews CE, Freedson PS, Hebert JR, Stanek EJ 3rd, Merriam PA, Rosal MC, Ebbeling CB, Ockene IS (2001b) Seasonal variation in household, occupational, and leisure time physical activity: longitudinal analyses from the seasonal variation of blood cholesterol study. Am J Epidemiol 153:172–183

Matthews CE, Moore SC, George SM, Sampson J, Bowles HR (2012) Improving self-reports of active and sedentary behaviors in large epidemiologic studies. Exerc Sport Sci Rev 40:118–126

McCormack GR, Friedenreich C, Shiell A, Giles-Corti B, Doyle-Baker PK (2010) Sex- and age-specific seasonal variations in physical activity among adults. J Epidemiol Community Health 64:1010–1016

McKenzie TL, Sallis JF, Nader PR, Patterson TL, Elder JP, Berry CC, Rupp JW, Atkins CJ, Buono MJ, Nelson JA (1991) BEACHES: an observational system for assessing children's eating and physical activity behaviors and associated events. J Appl Behav Anal 24:141–151

McKenzie TL, Marshall SJ, Sallis JF, Conway TL (2000) Leisure-time physical activity in school environments: an observational study using SOPLAY. Prev Med 30:70–77

McTiernan A (2008) Mechanisms linking physical activity with cancer. Nat Rev Cancer 8:205–211

McTiernan A (2011) Physical activity, dietary, calorie restriction, and cancer. Springer, New York

Merchant AT, Dehghan M, Akhtar-Danesh N (2007) Seasonal variation in leisure-time physical activity among Canadians. Can J Public Health 98:203–208

Merrill RM, Shields EC, White GL Jr, Druce D (2005) Climate conditions and physical activity in the United States. Am J Health Behav 29:371–381

Mueller C, Winter C, Rosenbaum D (2010) Current objective techniques for physical activity assessment in comparison with subjective methods. Deutsche Zeitschrift fuer Sportmedizin 61:11–18

Nader PR, Bradley RH, Houts RM, McRitchie SL, O'Brien M (2008) Moderate-to-vigorous physical activity from ages 9 to 15 years. JAMA 300:295–305

Neilson HK, Robson PJ, Friedenreich CM, Csizmadi I (2008) Estimating activity energy expenditure: how valid are physical activity questionnaires? Am J Clin Nutr 87:279–291

Nelson ME, Rejeski WJ, Blair SN, Duncan PW, Judge JO, King AC, Macera CA, Castaneda-Sceppa C (2007) Physical activity and public health in older adults: recommendation from the American College of Sports Medicine and the American Heart Association. Med Sci Sports Exerc 39:1435–1445

Ng SW, Popkin BM (2012) Time use and physical activity: a shift away from movement across the globe. Obes Rev 13:659–680

Nieman DC, Austin MD, Chilcote SM, Benezra L (2005) Validation of a new handheld device for measuring resting metabolic rate and oxygen consumption in children. Int J Sport Nutr Exerc Metab 15:186–194

Nilsson A, Anderssen SA, Andersen LB, Froberg K, Riddoch C, Sardinha LB, Ekelund U (2009) Between- and within-day variability in physical activity and inactivity in 9- and 15-year-old European children. Scand J Med Sci Sports 19:10–18

Norman A, Bellocco R, Bergstrom A, Wolk A (2001) Validity and reproducibility of self-reported total physical activity-differences by relative weight. Int J Obes Relat Metab Disord 25:682–688

Oja P, Bull FC, Fogelholm M, Martin BW (2010) Physical activity recommendations for health: what should Europe do? BMC Public Health 10:10. doi:10.1186/1471-2458-10-10

Opper E, Worth A, Wagner M, Bos K (2007) The module "Motorik" in the German Health Interview and Examination Survey for Children and Adolescents (KiGGS). Motor Fitness and physical activity of children and young people. Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz 50:879–888

OSG (1996) U.S. Department of Health and Human Services. Physical Activity and Health: A Report of the Surgeon General. Atlanta, GA: Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion

Owen N, Healy GN, Matthews CE, Dunstan DW (2010) Too much sitting: the population health science of sedentary behavior. Exerc Sport Sci Rev 38:105–113

Paffenbarger RS Jr, Blair SN, Lee IM (2001) A history of physical activity, cardiovascular health and longevity: the scientific contributions of Jeremy N Morris, DSc, DPH, FRCP. Int J Epidemiol 30:1184–1192

Parascandola M, Weed DL, Dasgupta A (2006) Two Surgeon General's reports on smoking and cancer: a historical investigation of the practice of causal inference. Emerg Themes Epidemiol 3:1. doi:10.1186/1742-7622-3-1

Pate RR (1993) Physical activity assessment in children and adolescents. Crit Rev Food Sci Nutr 33:321–326

Pate RR, Pratt M, Blair SN, Haskell WL, Macera CA, Bouchard C, Buchner D, Ettinger W, Heath GW, King AC, Kriska A, Leon AS, Marcus BH, Morris J, Paffenbarger RS, Patrick K, Pollock ML, Rippe JM, Sallis J, Wilmore JH (1995) Physical activity and public health. A recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine. JAMA 273:402–407

Pate RR, Almeida MJ, McIver KL, Pfeiffer KA, Dowda M (2006) Validation and calibration of an accelerometer in preschool children. Obesity (Silver Spring) 14:2000–2006

Patel RR, O'Neill JR, Lobelo F (2008) The evolving definition of "sedentary". Exerc Sport Sci Rev 36:173–178

Patel AV, Feigelson HS, Talbot JT, McCullough ML, Rodriguez C, Patel RC, Thun MJ, Calle EE (2008) The role of body weight in the relationship between physical activity and endometrial cancer: results from a large cohort of US women. Int J Cancer 123:1877–1882

Peters TM, Moore SC, Gierach GL, Wareham NJ, Ekelund U, Hollenbeck AR, Schatzkin A, Leitzmann MF (2009) Intensity and timing of physical activity in relation to postmenopausal breast cancer risk: the prospective NIH-AARP diet and health study. BMC Cancer 9:349. doi:10.1186/1471-2407-9-349

Pischon T, Hankinson SE, Hotamisligil GS, Rifai N, Rimm EB (2003) Leisure-time physical activity and reduced plasma levels of obesity-related inflammatory markers. Obes Res 11:1055–1064

Pollock ML, Gaesser GA, Butcher JD, Despres JP, Dishman RK, Franklin BA, Garber CE (1998) The recommended quantity and quality of exercise for developing and maintaining cardiorespiratory and muscular fitness, and flexibility in healthy adults. Med Sci Sports Exerc 30:975–991

President's Council on Physical Fitness and Sports (2008) Promoting physical activity using technology. Research Digest Ser 9(3):1–8

Proper KI, Singh AS, van Mechelen W, Chinapaw MJ (2011) Sedentary behaviors and health outcomes among adults: a systematic review of prospective studies. Am J Prev Med 40: 174–182

Quadrilatero J, Hoffman-Goetz L (2003) Physical activity and colon cancer. A systematic review of potential mechanisms. J Sports Med Phys Fit 43:121–138

Richardson MT, Ainsworth BE, Jacobs DR, Leon AS (2001) Validation of the Stanford 7-day recall to assess habitual physical activity. Ann Epidemiol 11:145–153

Ridley K, Dollman J, Olds T (2001) Development and validation of a computer delivered physical activity questionnaire (CDPAQ) for children. Pediatr Exerc Sci 13:35–46

Ridley K, Ainsworth BE, Olds TS (2008) Development of a compendium of energy expenditures for youth. Int J Behav Nutr Phys Act 5:45

Ridley K, Olds T, Hands B, Larkin D, Parker H (2009) Intra-individual variation in children's physical activity patterns: implications for measurement. J Sci Med Sport 12:568–572

Rockhill B, Willett WC, Manson JE, Leitzmann MF, Stampfer MJ, Hunter DJ, Colditz GA (2001) Physical activity and mortality: a prospective study among women. Am J Public Health 91: 578–583

Rodriguez BL, Curb JD, Burchfiel CM, Abbott RD, Petrovitch H, Masaki K, Chiu D (1994) Physical activity and 23-year incidence of coronary heart disease morbidity and mortality among middle-aged men. The Honolulu Heart Program. Circulation 89:2540–2544

Roger VL, Go AS, Lloyd-Jones DM, Benjamin EJ, Berry JD, Borden WB, Bravata DM, Dai S, Ford ES, Fox CS, Fullerton HJ, Gillespie C, Hailpern SM, Heit JA, Howard VJ, Kissela BM, Kittner SJ, Lackland DT, Lichtman JH, Lisabeth LD, Makuc DM, Marcus GM, Marelli A, Matchar DB, Moy CS, Mozaffarian D, Mussolino ME, Nichol G, Paynter NP, Soliman EZ, Sorlie PD, Sotoodehnia N, Turan TN, Virani SS, Wong ND, Woo D, Turner MB (2012) Heart disease and stroke statistics–2012 update: a report from the American Heart Association. Circulation 125:e2–e220

Rothman KJ, Greenland S (2005) Causation and causal inference in epidemiology. Am J Public Health 95(Suppl 1):S144–S150

Rowe PJ, Schuldheisz JM, vanderMars H (1997) Validation of SOFIT for measuring physical activity of first- to eighth-grade students. Pediatr Exerc Sci 9:136–149

Sabia S, Dugravot A, Kivimaki M, Brunner E, Shipley MJ, Singh-Manoux A (2012) Effect of intensity and type of physical activity on mortality: results from the Whitehall II cohort study. Am J Public Health 102:698–704

Sallis JF, Saelens BE (2000) Assessment of physical activity by self-report: status, limitations, and future directions. Res Q Exerc Sport 71:S1–S14

Sallis JF, Haskell WL, Wood PD, Fortmann SP, Rogers T, Blair SN, Paffenbarger RS Jr (1985) Physical activity assessment methodology in the Five-City Project. Am J Epidemiol 121: 91–106

Sallis JF, Prochaska JJ, Taylor WC (2000) A review of correlates of physical activity of children and adolescents. Med Sci Sports Exerc 32:963–975

Saris WH, Blair SN, van Baak MA, Eaton SB, Davies PS, Di Pietro L, Fogelholm M, Rissanen A, Schoeller D, Swinburn B, Tremblay A, Westerterp KR, Wyatt H (2003) How much physical activity is enough to prevent unhealthy weight gain? Outcome of the IASO 1st Stock Conference and consensus statement. Obes Rev 4:101–114

Schatzkin A, Subar AF, Moore S, Park Y, Potischman N, Thompson FE, Leitzmann M, Hollenbeck A, Morrissey KG, Kipnis V (2009) Observational epidemiologic studies of nutrition and cancer: the next generation (with better observation). Cancer Epidemiol Biomark Prev 18:1026–1032

Schmidt MD, Freedson PS, Chasan-Taber L (2003) Estimating physical activity using the CSA accelerometer and a physical activity log. Med Sci Sports Exerc 35:1605–1611

Schmitz KH, Courneya KS, Matthews C, Demark-Wahnefried W, Galvao DA, Pinto BM, Irwin ML, Wolin KY, Segal RJ, Lucia A, Schneider CM, von Gruenigen VE, Schwartz AL (2010) American College of Sports Medicine roundtable on exercise guidelines for cancer survivors. Med Sci Sports Exerc 42:1409–1426

Schneider PL, Crouter SE, Lukajic O, Bassett DR Jr (2003) Accuracy and reliability of 10 pedometers for measuring steps over a 400-m walk. Med Sci Sports Exerc 35:1779–1784

Schoeller DA, Webb P (1984) Five-day comparison of the doubly labeled water method with respiratory gas exchange. Am J Clin Nutr 40:153–158

Schoeller DA, Ravussin E, Schutz Y, Acheson KJ, Baertschi P, Jequier E (1986) Energy expenditure by doubly labeled water: validation in humans and proposed calculation. Am J Physiol 250:R823–R830

Schoenborn CA, Stommel M (2011) Adherence to the 2008 adult physical activity guidelines and mortality risk. Am J Prev Med 40:514–521

Sharma S, Chuang RJ, Skala K, Atteberry H (2011) Measuring physical activity in preschoolers: reliability and validity of The System for Observing Fitness Instruction Time for Preschoolers (SOFIT-P). Meas Phys Educ Exerc Sci 15:257–273

Sherwood NE, Jeffery RW (2000) The behavioral determinants of exercise: implications for physical activity interventions. Annu Rev Nutr 20:21–44

Sirard JR, Pate RR (2001) Physical activity assessment in children and adolescents. Sports Med 31:439–454

Sirard JR, Trost SG, Pfeiffer KA, Dowda M, Pate RR (2005) Calibration and evaluation of an objective measure of physical activity in preschool children. J Phys Act Health 3:345–357

Slattery ML, Potter JD (2002) Physical activity and colon cancer: confounding or interaction? Med Sci Sports Exerc 34:913–919

Spurr GB, Prentice AM, Murgatroyd PR, Goldberg GR, Reina JC, Christman NT (1988) Energy expenditure from minute-by-minute heart-rate recording: comparison with indirect calorimetry. Am J Clin Nutr 48:552–559

Starling DS (2002) Use of doubly labeled water and indirect calorimetry to assess physical activity. In: Welk GJ (ed) Physical activity assessments for health-related research. Human Kinetics, Inc., Champaign, pp 197–209

Steele BG, Belza B, Cain K, Warms C, Coppersmith J, Howard J (2003) Bodies in motion: monitoring daily activity and exercise with motion sensors in people with chronic pulmonary disease. J Rehabil Res Dev 40:45–58

Sternfeld B, Dugan S (2011) Physical activity and health during the menopausal transition. Obstet Gynecol Clin North Am 38:537–566

Strath SJ, Swartz AM, Bassett DR Jr, O'Brien WL, King GA, Ainsworth BE (2000) Evaluation of heart rate as a method for assessing moderate intensity physical activity. Med Sci Sports Exerc 32:S465–S470

Strath SJ, Brage S, Ekelund U (2005) Integration of physiological and accelerometer data to improve physical activity assessment. Med Sci Sports Exerc 37:S563–S571

Sui X, LaMonte MJ, Laditka JN, Hardin JW, Chase N, Hooker SP, Blair SN (2007) Cardiorespiratory fitness and adiposity as mortality predictors in older adults. JAMA 298:2507–2516

Susser M (1991) What is a cause and how do we know one? A grammar for pragmatic epidemiology. Am J Epidemiol 133:635–648

Telford A, Salmon J, Jolley D, Crawford D (2004) Reliability and validity of physical activity questionnaires for children: The Children's Leisure Activities Study Survey (CLASS). Pediatr Exerc Sci 16:64–78

Thorp AA, Owen N, Neuhaus M, Dunstan DW (2011) Sedentary behaviors and subsequent health outcomes in adults a systematic review of longitudinal studies, 1996–2011. Am J Prev Med 41:207–215

Trost G (2007) State of the art reviews: measurement of physical activity in children and adolescents. Am J Lifstyle Med. doi:10.1177/1559827607301686

Trost SG, Owen N, Bauman AE, Sallis JF, Brown W (2002) Correlates of adults' participation in physical activity: review and update. Med Sci Sports Exerc 34:1996–2001

Trost SG, McIver KL, Pate RR (2005) Conducting accelerometer-based activity assessments in field-based research. Med Sci Sports Exerc 37:S531–S543

Tudor-Locke CE, Myers AM (2001) Challenges and opportunities for measuring physical activity in sedentary adults. Sports Med 31:91–100

Tudor-Locke C, Bassett DR Jr (2004) How many steps/day are enough? Preliminary pedometer indices for public health. Sports Med 34:1–8

Tudor-Locke C, Williams JE, Reis JP, Pluto D (2002) Utility of pedometers for assessing physical activity: convergent validity. Sports Med 32:795–808

Tudor-Locke C, Washington TL, Ainsworth BE, Troiano RP (2009) Linking the American Time Use Survey (ATUS) and the compendium of physical activities: methods and rationale. J Phys Act Health 6:347–353

U.S. Department of Health and Human Services (1996) Physical activity and health: a report of the Surgeon General. U.S. Department of Health and Human Services/Centers for Disease Control and Prevention/National Center for Chronic Disease Prevention and Health Promotion, Atlanta

U.S. Department of Health and Human Services (1998) Physical activity and health: a report of the Surgeon General. U.S. Department of Health and Human Services/Centers for Disease Control and Prevention/National Center for Chronic Disease Prevention and Health Promotion, Atlanta

U.S. Department of Health and Human Services (2008) Physical activity guidelines for Americans. Washington, DC. http://www.health.gov/paguidelines/pdf/paguide.pdf. Accessed 8 Aug 2012

U.S. Department of Health and Human Services, U.S. Department of Agriculture (2000) Nutrition and your health: dietary guidelines for Americans, 5th edn. U.S. Government Printing Office, Washington, DC

van Cauwenberghe E, Labarque V, Trost SG, de Bourdeaudhuij I, Cardon G (2011) Calibration and comparison of accelerometer cut points in preschool children. Int J Pediatr Obes 6:e582–e589

van den Berg MH, Schoones JW, Vliet Vlieland TP (2007) Internet-based physical activity interventions: a systematic review of the literature. J Med Internet Res 9:e26

Vanhees L, Lefevre J, Philippaerts R, Martens M, Huygens W, Troosters T, Beunen G (2005) How to assess physical activity? How to assess physical fitness? Eur J Cardiovasc Prev Rehabil 12:102–114

Victora CG, Habicht JP, Bryce J (2004) Evidence-based public health: moving beyond randomized trials. Am J Public Health 94:400–405

Voorrips LE, Ravelli AC, Dongelmans PC, Deurenberg P, Van Staveren WA (1991) A physical activity questionnaire for the elderly. Med Sci Sports Exerc 23:974–979

Wall MI, Carlson SA, Stein AD, Lee SM, Fulton JE (2011) Trends by age in youth physical activity: Youth Media Campaign Longitudinal Survey. Med Sci Sports Exerc 43:2140–2147

Walsh NP, Gleeson M, Shephard RJ, Woods JA, Bishop NC, Fleshner M, Green C, Pedersen BK, Hoffman-Goetz L, Rogers CJ, Northoff H, Abbasi A, Simon P (2011) Position statement. Part one: Immune function and exercise. Exerc Immunol Rev 17:6–63

Wannamethee SG, Shaper AG, Walker M (1998) Changes in physical activity, mortality, and incidence of coronary heart disease in older men. Lancet 351(9116):1603–1608

Ward DS, Evenson KR, Vaughn A, Rodgers AB, Troiano RP (2005) Accelerometer use in physical activity: best practices and research recommendations. Med Sci Sports Exerc 37:S582–S588

Wareham NJ, Hennings SJ, Prentice AM, Day NE (1997) Feasibility of heart-rate monitoring to estimate total level and pattern of energy expenditure in a population-based epidemiological study: the Ely Young Cohort Feasibility Study 1994–5. Br J Nutr 78:889–900

Wareham NJ, Jakes RW, Rennie KL, Mitchell J, Hennings S, Day NE (2002) Validity and repeatability of the EPIC-Norfolk Physical Activity Questionnaire. Int J Epidemiol 31: 168–174

Warren JM, Ekelund U, Besson H, Mezzani A, Geladas N, Vanhees L (2010) Assessment of physical activity – a review of methodologies with reference to epidemiological research: a report of the exercise physiology section of the European Association of Cardiovascular Prevention and Rehabilitation. Eur J Cardiovasc Prev Rehabil 17:127–139

Washburn RA (2000) Assessment of physical activity in older adults. Res Q Exerc Sport 71:S79–S88

Wei M, Gibbons LW, Kampert JB, Nichaman MZ, Blair SN (2000) Low cardiorespiratory fitness and physical inactivity as predictors of mortality in men with type 2 diabetes. Ann Intern Med 132:605–611

Weir JB (1949) New methods for calculating metabolic rate with special reference to protein metabolism. J Physiol 109:1–9

Welk GJ, Blair SN, Wood K, Jones S, Thompson RW (2000) A comparative evaluation of three accelerometry-based physical activity monitors. Med Sci Sports Exerc 32:S489–S497

Wen CP, Wai JP, Tsai MK, Yang YC, Cheng TY, Lee MC, Chan HT, Tsao CK, Tsai SP, Wu X (2011) Minimum amount of physical activity for reduced mortality and extended life expectancy: a prospective cohort study. Lancet 378:1244–1253

Westerterp KR (2001) Pattern and intensity of physical activity. Nature 410:539

Willett W (1998) Nutritional epidemiology, 2nd edn. Oxford University Press, New York

Winters-Hart CS, Brach JS, Storti KL, Trauth JM, Kriska AM (2004) Validity of a questionnaire to assess historical physical activity in older women. Med Sci Sports Exerc 36:2082–2087

Wolf AM, Hunter DJ, Colditz GA, Manson JE, Stampfer MJ, Corsano KA, Rosner B, Kriska A, Willett WC (1994) Reproducibility and validity of a self-administered physical activity questionnaire. Int J Epidemiol 23:991–999

Wolin KY, Yan Y, Colditz GA, Lee IM (2009) Physical activity and colon cancer prevention: a meta-analysis. Br J Cancer 100:611–616

World Health Organization (2005) HEPA Europe (European network for the promotion of health-enhancing physical activity). http://www.euro.who.int/__data/assets/pdf_file/0019/101692/Brochure-on-health-enhancing-physical-activity-Eng.pdf. Accessed 25 Sept 2012

World Health Organization (2011) Global recommendations on physical activity for health. http://whqlibdoc.who.int/publications/2010/9789241599979_eng.pdf. Accessed 2 July 2012

World Health Organization (2012a) Global physical activity surveillance. http://www.who.int/chp/steps/GPAQ/en/index.html. Accessed 10 Aug 2012

World Health Organization (2012b) Chronic diseases and health promotion. Global physical activity surveillance. http://www.who.int/chp/steps/GPAQ/en/index.html. Accessed 26 Sept 2012

Zhang K, Werner P, Sun M, Pi-Sunyer FX, Boozer CN (2003) Measurement of human daily physical activity. Obes Res 11:33–40

Zhu WM, Hasegawa-Johnson M, Roth D, Kantor A, Gao Y, Gandhi MA, Park Y, Yang L (2006) Validation of an E-diary system for assessing physical activities. Med Sci Sports Exerc 38:S102–S103

Radiation Epidemiology

50

Hajo Zeeb, Hiltrud Merzenich, Henryk Wicke, and Maria Blettner

Contents

H. Zeeb (✉)
Department of Prevention and Evaluation, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

Faculty of Human and Health Sciences, University of Bremen, Bremen, Germany

H. Merzenich • H. Wicke
Institute for Medical Biostatistics, Epidemiology and Informatics, University Medical Center, Johannes Gutenberg University Mainz, Mainz, Germany

M. Blettner
Institute for Medical Biostatistics, Epidemiology and Informatics, Johannes Gutenberg University, Mainz, Germany

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_68,

50.1 Introduction

Radiation epidemiology aims at investigating the health effects of radiation exposure due to different types of radiation. It is one of the best established specialized fields of epidemiology. The discovery of ionizing radiation in 1895 led to a wealth of subsequent scientific studies on physical, biological, and clinical radiation effects. Medical scientists and epidemiologists have continued to show intense interest in exploring health effects of radiation exposure in environmental, medical, military, and occupational settings, and the large studies of radiation-exposed populations belong to the milestones of international epidemiological research.

Under the label of radiation epidemiology, research on different types of radiation and their health effects is being conducted. The effects of radiation can roughly be categorized to be either ionizing or non-ionizing. Non-ionizing radiation includes electric and magnetic fields and ultraviolet radiation. Ionizing radiation, the topic of this chapter, probably belongs to the best studied human carcinogens, comparable only to tobacco smoke. Both observational epidemiology and experimental research have provided conclusive evidence on the carcinogenicity of ionizing radiation (IARC 2000). However, other effects of ionizing radiation are also well understood and technologically utilized, perhaps most notably the cell-killing properties of ionizing radiation in cancer therapy.

In radiation epidemiology, one large study continues to be a cornerstone for risk estimation and a benchmark for other studies in the field: the Life Span cohort Study (LSS) of the atomic bomb survivors in Hiroshima and Nagasaki in 1945. The study is continuously being updated and now covers a follow-up period of more than 50 years (Ozasa et al. 2012). The influence of this one study sets the field apart from other areas of epidemiology where no such reference exists. Beyond the LSS, an ever-increasing number of epidemiological studies with improved exposure assessment and study methodology contribute to the current knowledge about health effects of ionizing radiation.

Developments in radiation epidemiology have not only been important for an improvement in the understanding of health effects and the subsequent adaptation of radiation protection standards but also for methodological advances in the planning, conduct, and analysis of epidemiological studies. Hence, radiation epidemiology is both an important contributor to the historical scientific development of epidemiology and to ongoing research, methods development, and risk communication. This chapter gives an overview of radiation sources and exposures and then focuses on exposure assessment and examples of studies in the field. The importance of radiation epidemiology for radiation protection as well as current developments in research and open questions are highlighted.

50.2 Sources of Radiation Exposure

There are various natural and man-made sources of ionizing radiation, and virtually all of these sources have been studied with regard to their health effects for humans. A major part of population exposure to ionizing radiation comes from natural sources. Natural radiation exposure is composed of internal and external components.

Via inhaled air and nutrition, human beings have always been exposed to natural radioactive substances. The inhalation of radon gas generally is the largest natural source of radiation contributing to the exposure of the general population. Radon (radon-222) has a half-life of 3.8 days and is the most stable isotope in the radium and uranium decay chain. Radon exposure originates from rocks and soils and varies geographically depending on the local content of radionuclides like uranium and thorium. Indoor exposure to radon gas depends on radionuclide concentrations in soil, shielding and isolation, as well as on building materials (WHO 2009).

Furthermore, radioactive isotopes in food and water lead to internal radiation exposure. Exposures from eating and drinking water are due in part to the uranium and thorium series. Radioactive carbon (carbon-14) is present in all organic matter, accumulates in the food chain, and contributes to the internal background dose from ionizing radiation.

Terrestrial radiation originates from natural radioactive substances which exist in the soils and rock layers of the earth's crust. Rocks and soils are also important raw materials for building materials such as bricks and concrete. Hence, the radionuclides contained in these materials also contribute to external radiation exposure of residents in buildings.

Cosmic rays are radiation that travels through interstellar space. Known sources of this radiation are, among others, the sun and supernovae (exploding stars). The earth's magnetic field provides some protection from cosmic radiation. On the earth's surface, cosmic rays lead to effective doses of around 0.3 millisievert (mSv) per year (for explanation of dose quantities, see Sect. 50.3). The exposure increases with increasing altitude, notably during air travel (Bartlett 2004).

The per capita effective annual dose from all natural radiation sources combined is estimated to be around 2.4 mSv (UNSCEAR 2008). Of this amount, about 1.1 mSv per year comes from radon and its decay products. However, the individual exposure to natural radiation varies considerably due to large differences in environmental concentrations and in population habits (Hendry et al. 2009).

In addition to natural sources, man-made sources contribute to the individual annual dose with an estimated worldwide average of about 0.6 mSv. In many developed countries, these doses are much higher. In Germany, for example, average doses from man-made sources amount to about 2.0 mSv per year. Medical uses of radiation are largely determined by the degree of development of the respective health-care system and contribute the greater part of these exposures, while occupational sources, nuclear power generation, and fallout from nuclear weapons tests generally make comparatively minor contributions. As an example, Fig. 50.1 shows the relative contribution of different radiation sources in Germany (Federal Office for Radiation Protection 2012).

In the field of diagnostic radiology, a rapid increase of computed tomography (CT) use was observed in many countries during the past 20–30 years. The estimated annual number of CT examinations in the United States rose approximately

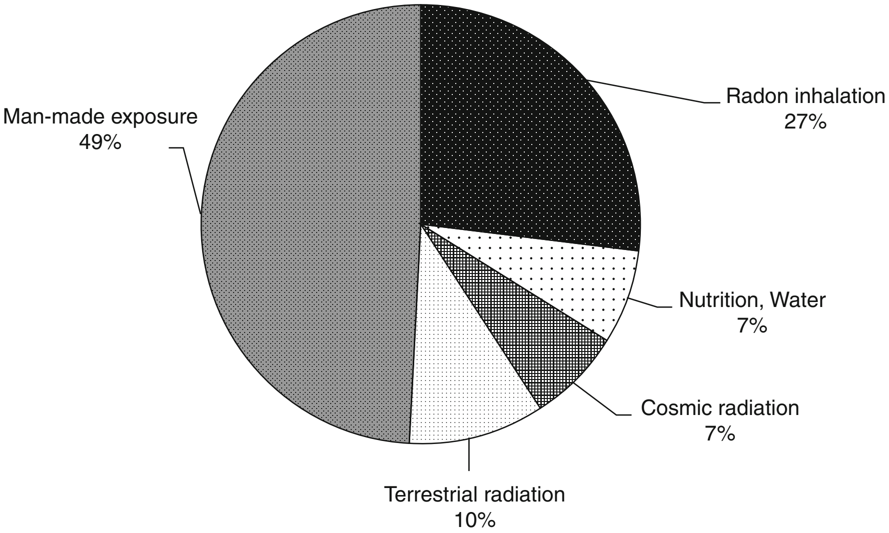


Fig. 50.1 Sources of ionizing radiation in Germany (proportion of total effective dose in %)

Table 50.1 Average effective dose levels from medical procedures compared to the doses of atomic bomb survivors

Medical procedure	Effective dose in mSv[a]
Chest X-ray	0.01
Dental X-ray	0.01
Mammography	0.2
CT chest	10
CT abdomen and pelvis	15
Radiotherapy (average dose to target organ)	30,000–70,000
Atomic bomb survivors (mean dose)	∼200

[a] *mSv* millisievert, generally equal to *mGy* milligray in the examples given here; see Sect. 50.3 for dose quantities

sevenfold from 2.8 million in 1981 to 20 million in 1995 (Brenner et al. 2001). The CT radiation doses are relatively high compared to most conventional X-ray examinations. Thus, CT is a major contributor to the collective diagnostic dose of the population in developed countries. CT doses are of particular concern in pediatric radiological imaging. Compared to adults, children are at higher risk for developing cancer caused by ionizing radiation mainly due to an increased radiosensitivity and a longer life span after exposure (Brenner et al. 2001). To date, ionizing radiation is the only established environmental risk factor for childhood leukemia (Greaves 2006; Lightfoot and Roman 2004).

Table 50.1 shows the effective dose levels from medical procedures in relation to the mean dose of atomic bomb survivors. Note that doses vary considerably depending on technical and patient characteristics.

Therapeutic uses of ionizing radiation differ from diagnostic radiology procedures both in purpose and in the employed doses. Radiotherapy is an important treatment option for many malignant diseases. In most cases, patients receive high doses of radiation in the order of 40–60 gray (Gy) to the target region. Studies of patients treated with radiation provide important information about the carcinogenicity of radiation, on the effects of different radiation types, and on the role of sex, age, or genetic susceptibility (National Research Council 2006). Large cohorts of radiation-treated patients who have been followed for long periods are available and allow the assessment of the risk of a second primary cancer. An example is the British Childhood Cancer Survivor Study (Reulen et al. 2011) (see also Sect. 50.5).

Ionizing radiation was also used for the treatment of certain benign diseases, for example, benign breast diseases, peptic ulcer, and benign thyroid diseases; there are very restricted uses for ionizing radiation in this context nowadays. Even among children, radiotherapy was used to treat diseases such as fungal scalp infections (tinea capitis), enlarged thymus gland, or benign head and neck diseases. In a US cohort of 601 women treated with radiotherapy for acute postpartum mastitis (APM) between 1940 and 1957, the average dose to the breasts was 3.8 Gy. In comparison to non-irradiated control groups, a more than threefold relative risk (RR) for breast cancer of the irradiated breasts was found (Shore et al. 1986).

Occupational exposures account for only about 2% of the man-made exposure overall (National Research Council 2006). However, for specific groups of occupationally exposed workers, these exposures can be substantial. Persons with occupations involving the medical use of radiation were among the first groups where adverse health effects could be demonstrated (March 1944). Other groups include uranium miners, such as the large group of mine workers in the former German Democratic Republic who were highly exposed to radioactive radon (Walsh et al. 2010). Employees in the nuclear power production are exposed to low protracted doses of radiation (Cardis et al. 2005).

Nuclear weapons testing was (with few exceptions) conducted between 1945 and 1980. These tests led to worldwide dispersion of radioactive material in the atmosphere (nuclear fallout) and subsequent radiation exposure. The worldwide average annual effective dose reached a peak of 110 microsievert (μSv) per person in 1963 and has since decreased to about 5 μSv/year. Furthermore, several accidents have led to releases of radioactivity to the environment, among them the Kyshtym accident in 1957 (Russia), the Windscale accident in 1957 (UK), the Accident at Three Mile Island in 1979 (USA), the Chernobyl accident in 1986 (Ukraine) (UNSCEAR 2008), and the Fukushima accident in 2011 (Japan). Children and adolescents exposed to radioactive iodine from the Chernobyl plant showed a dose-related increase in thyroid cancer occurrence (Cardis and Hatch 2011).

In 1945, at the time of the detonation of the atomic bombs, about 500,000 people lived in Hiroshima and Nagasaki. About 200,000 died immediately due to superficial burns, acute radiation syndrome, or injuries. In 1950, the National Census of A-Bomb-Survivors registered about 280,000 victims affected by the bombings. Japanese and American scientists recruited a cohort of 120,000 persons from among the survivors and set up what is now known as the Life Span Study (LSS). This cohort study has become the principal basis for estimating radiation risk of cancer and other diseases because it comprises a wide range of radiation exposures and a relatively accurate dosimetry with estimated radiation doses to numerous organs and tissues (National Research Council 2006) (see also Sects. 50.3 and 50.5).

50.3 Radiation Exposure: Doses and Measurement

50.3.1 Basic Principles of Radiation Physics

Radiation is energy or, equivalently, matter that travels through space and time. It can be roughly categorized into particle radiation and electromagnetic waves – both of which will be described in more detail below. Along its way, radiation can physically interact with the irradiated matter. Depending on the energy content of the radiation, it may either be of the *ionizing* or of the *non-ionizing* variety. The former possesses an energy content larger than the electronic binding energy, therefore possibly altering the electronic configuration of atoms and molecules. The ionization process leads to the emission of a shell electron from the Coulomb field of the atomic nucleus and leaves behind charged atomic or molecular ions.

Ionizing radiation can be many different physical entities, that is, charged energetic particles (e.g., protons, muons, pions, alpha particles, or heavier clusters), neutrons, or high energy, that is, short-wavelength electromagnetic radiation. With non-ionizing radiation, one usually refers to comparatively low-energy, long-wavelength electromagnetic radiation.

Because the chemical bonding pattern between atoms or molecules is determined by their electronic structures, ionization processes can break chemical bonds. Damage of the organic molecule deoxyribonucleic acid (DNA) by ionizing radiation can eventually lead to the development of cancer in the affected individual.

There are stable and instable atomic nuclei. The latter are also called *radioactive nuclei* or *radionuclides*, for which their exact time of decay cannot be predicted as radioactive decay is a stochastic process that, in almost all cases, cannot be influenced from the outside. This is reflected by the concept of *half-life*, which is the time span in which the activity of a specific radioactive substance decreases to 50% of its initial value. The half-life $T_{1/2}$, the mean lifetime τ, and the so-called decay constant λ, which pertain to the probability of a radioactive nucleus to change its state within a given time interval, are related via

$$T_{1/2} = \ln(2/\lambda) = \tau \ln 2.$$

The SI unit (French: Système international d'unités, international system of units) for the activity of a radioactive substance, that is, the number of radioactive transformations per time, is the reciprocal second (s^{-1}) called *becquerel* (Bq). An obsolete, historical unit for the activity is the *curie* (Ci).

If a radioactive atomic nucleus decays in order to attain an energetically more favorable configuration, energy is emitted, in many cases, in the form of *alpha*, *beta*, or *gamma* radiation.

Alpha particles are simply the doubly positively charged atomic nuclei of the element helium, so they consist of two protons and two neutrons. They are predominantly produced by heavier instable nuclei. Two prominent examples of alpha-emitting radionuclides in the context of epidemiology are the isotopes of radium (Martland and Humphries 1929) and radon (Grosche et al. 2006).

Beta radiation consists of elementary particles of negatively charged electrons (β^- radiation) or positively charged positrons (β^+ radiation).

Gamma decay is the emission of nuclear excitation energy in the form of electromagnetic radiation. Here, the mass number, that is, the nucleus' total number of neutrons and protons, remains unchanged, while the energy and actual mass of the nucleus decreases. A prominent example is the gamma decay of the metastable state of barium-137 to that isotope's ground state following the beta-decay of cesium-137 to metastable barium-137. Cesium-137 is a radiologically relevant radionuclide, for example, in nuclear accidents like those in Chernobyl or Fukushima. Hence, the respective emission of 662 kiloelectronvolt (keV) gamma radiation by metastable barium-137 is routinely used for the physical detection of cesium-137 using gamma spectrometry.

Gamma radiation can be regarded both as electromagnetic waves or as another type of elementary particles (photons). In that, it is only distinguishable from other types of electromagnetic radiation, for example, visible light, by its energy content or, equivalently, its (short) wavelength which enables ionization of atoms or molecules. This does not only include gamma radiation from radioactive decay but also the X-rays well known from medical applications and also short-wavelength parts of the ultraviolet spectrum (ultraviolet light does not derive its particular skin cancerogenicity from this fact, though). All other electromagnetic radiation, including the infrared, visible, and radio wave parts of the electromagnetic spectrum, belong to the non-ionizing radiation.

Ionizing radiation can be categorized by the way it causes the ionization of the irradiated target. Charged particles like alpha particles, beta particles, or protons belong to the directly ionizing radiation because these particles are mainly responsible for the ionization of the target. Indirectly ionizing radiation includes the uncharged photons and neutrons that produce charged secondary particles, which are *mainly* responsible for the ionization of the target matter. This does not mean, though, that indirectly ionizing radiation like neutron radiation is less dangerous than directly ionizing radiation (see *relative biological effectiveness* in the following section).

The shielding of ionizing radiation heavily depends on the exact type of radiation. Alpha particles, on the one hand, are typically stopped in air after a few centimeters, depending on their energy, and can be effectively shielded by a sheet of paper. Hence, they derive their radiological relevance mainly from situations where they have been incorporated through ingestion or inhalation. Beta particles, on the other hand, can travel some meters in air but are usually stopped by just a few millimeters of aluminum. For the shielding of photons, thick layers of lead or other absorbers with high atomic numbers are necessary. The situation is again different for neutrons, where substances like paraffin or water are most suitable for moderation or shielding because they are facilitating neutron-proton scattering (Grupen 2010).

Humans cannot sense ionizing radiation. For the purpose of surveillance and exposure limitation as well as for exposure assessment in epidemiological and other studies, devices to measure radiation are needed. The most simple gas-filled radiation detector is the ionization chamber, which, in principle, already allows some spectroscopic analyses. The fiber or pen dosimeter is a common ionization chamber for radiation protection and personal dose measurement. Proportional and Geiger-Müller counters are based on similar physical principles as the ionization chamber but, depending on their respective design, can be used as robust measurement devices for contaminations or ambient doses. In thermoluminescence dosimeters, used, for example, in the form of film badge dosimeters, the blackening of a photographic film is employed as a measure for the received radiation dose. Radiation measurement of environmental samples is nowadays often conducted using (liquid) scintillation counters or semiconductor detectors allowing high-resolution spectroscopy for the differentiation of the various radionuclides.

50.3.2 Dose Quantities and Dosimetry

Since radiation effects on humans are ultimately based on the absorption of radiation energy in tissue, the *energy dose D* can be regarded as a fundamental physical dose quantity. It is the absorbed energy per mass of the absorbing material. It has the unit *gray* (Gy), derived from the SI units for energy (joule, J) and mass (kilograms, kg):

$$1 \text{ Gy} = 1 \text{ J kg}^{-1}.$$

Other physical dose quantities are the *ion dose* and the *kerma*. The ion dose is the produced electrical charge per air mass in an air volume irradiated with ionizing radiation. It has the unit coulomb per kilogram (C kg^{-1}).

The word “kerma” is an acronym for “kinetic energy released per unit mass” (Grupen 2010) and refers to the ratio of the kinetic energy transferred to charged secondary particles from indirect ionizing radiation and the mass of the irradiated volume. The kerma also has the unit Gy.

The described physical dose quantities do not accurately describe the relative biological effectiveness of the various types of radiation. Alpha radiation, for example, belongs to the *densely ionizing radiation*. Its high ionization density entails less effective biological repair mechanisms. Alpha radiation thus has a higher biological effectiveness than, for example, the *sparsely ionizing* electrons, as the response or reaction of a biological system to these different types of radiation differs substantially. Hence, in radiation protection, special dose quantities are used

Table 50.2 List of radiation weighting factors (ICRP 2007)

Type of radiation	w_R
Photons	1
Electrons and muons	1
Protons and charged pions	2
Neutrons	5–20
Alpha particles, fission fragments, heavy nuclei	20

that are derived from the energy dose, but cannot be regarded as physical dose quantities. To differentiate these dose quantities from the energy dose, the unit *sievert* (Sv) with the same relation to SI units is being used:

$$1 \text{ Sv} = 1 \text{ J kg}^{-1}.$$

Thus, Gy and Sv are the same in terms of SI units, but Gy relates simply to an energy dose, whereas Sv includes a measure of biological reaction to the respective type of radiation not included in Gy. Further biological aspects of cell, tissue, and organ response to radiation need to be taken into account when aiming to assess radiation effects on the body.

With the development of a comprehensive system of radiation protection, this became important in the context of efforts to better quantify the risk of stochastic (cancer and genetic) effects. These efforts included the estimation of cancer risks for a number of organs and the concept of a weighted whole-body dose as summary value. Weights were based on the organ-specific cancer risk. The main idea was to create a single value with which different types of radiation and different distributions of radiation in the body could be compared. This resulted in the effective dose concept, details of which are explained in the following section.

Energy dose D and the specific radiation type can be linked via the equivalent dose H_R which is calculated from the energy dose D using

$$H_R = w_R \cdot D$$

with the radiation weighting factor w_R for the radiation type R. By means of this factor, the different relative biological effectiveness of the various radiation types can be accounted for (Table 50.2).

For the radiation weighting factor of neutrons, an energy-dependent set of equations is used (ICRP 2007). The typical range of neutron radiation weighting factors is about 5–20, depending on the energy range of neutrons which can vary substantially.

The organ dose H_T relates to an irradiated organ or tissue type T. It is calculated from the energy dose D_T over organ or tissue type T using

$$H_T = w_R \cdot D_T.$$

Table 50.3 List of tissue weighting factors (ICRP 2007)

Organ/tissue	w_T
Gonads	0.08
Red bone marrow	0.12
Colon	0.12
Lungs	0.12
Stomach	0.12
Bladder	0.04
Breasts	0.12
Liver	0.04
Esophagus	0.04
Thyroid	0.04
Skin	0.01
Bone surfaces	0.01
Salivary glands	0.01
Brain	0.01
Other	0.12

The effective dose E combines the (radiation weighted) organ doses with tissue weighting factors w_T, whose purpose is to represent the different radiation sensitivities of the organs (Table 50.3):

$$E = \sum_T w_T \cdot H_T .$$

The tissue weighting factors aim to take into account the different biological and physiological conditions and functionalities of the various organs, resulting in different susceptibilities for radiation-related detriment. As can be seen from Table 50.3, red bone marrow is estimated to be about three times as sensitive to radiation as, for example, the liver. Hence, the concept of the effective dose allows the calculation of weighted whole-body doses, which serve as a somewhat abstract measure (the effective dose cannot be directly measured) for the whole-body radiation-related detriment.

50.3.3 Assessment of Radiation Exposure in Epidemiological Studies

The validity and informative value of studies in radiation epidemiology depends on the accuracy of the dose assessment. As the type of radiation to be expected is usually known from the physical properties of the radiation source, appropriate dosimetric approaches can be selected purposefully. Various possibilities for quantifying the radiation exposure exist: direct measurement or (indirect) estimation of personal doses, personal- or population-based dose estimation or modeling based on environmental measurements, job-exposure matrices, or dosimetry systems (often developed in the context of large-scale epidemiological studies) combining different approaches.

Certainly, the most straightforward way to assess the radiation exposure of persons included in an epidemiological study is to measure it directly. Using personal dosimeters, one can compile individual person doses, thereby attaining relatively accurate values, which are mostly limited by the uncertainties associated with the physical measurement technique. Obviously, such direct measurements are rarely available as they generally pertain to controlled situations where the exposed persons have been given dosimeters to begin with, for instance nuclear industry workers (Vrijheid et al. 2007). Even then, as, for example, in the international study of nuclear industry workers (Thierry-Chef et al. 2007), changes over time and comparability of measurements from different facilities and countries pose challenges and need to be studied in detail.

Medical radiation exposure, be it for therapeutic or diagnostic purposes for patients or as occupational exposure of personnel, is an important topic of epidemiological research (Boice 1981; Hall and Brenner 2008). In the case of medical personnel, direct measurement of personal doses is certainly desirable and sometimes also possible (Häusler et al. 2009). Yet, epidemiologists often have to resort to more indirect dose surrogates as, for example, in the historical reconstruction of doses received by US radiological technologists included in a large cohort study (Simon et al. 2006). In this study, personal dose measurements available for some cohort members were combined with various, partly time-dependent circumstantial information like work histories, radiation protection practices, and other work conditions, based on measurements, literature data, or assumptions, to yield, after the application of a Monte Carlo simulation approach, individual, probabilistic personal dose estimates.

In the context of epidemiological studies, dose estimation of medical radiation exposures may, as a minimum, include the number of, for example, diagnostic X-ray or computed tomography examinations. If examination parameters are known, dose estimation may be conducted, possibly extended by individual organ dose estimates.

Ongoing research, especially in the area of computed tomography, aims at more detailed and precise dose estimates, including organ doses, based on a thorough analysis of various examination parameters, for example, technical parameters of the computed tomography system or physiological parameters of the patient. The development of several computer codes is ongoing to automate such increasingly complex dose assessments (Stamm and Nagel 2002; Kim et al. 2011).

Epidemiological studies dealing with uncontrolled situations, most prominently nuclear accidents, are another example where personal dosimetry is usually not available. In such situations, one usually has to rely on radioecological measurements of radioactivity in environmental samples, be it soils, locally produced food, or other potentially contaminated material. Following the Chernobyl accident in 1986, extensive measurements of the radionuclide cesium-137 in the environment were conducted (Drozdovitch et al. 2007). Another relevant radionuclide from the Chernobyl fallout was iodine-131, which is very important in the context of studies on Chernobyl-related thyroid cancer. The beta emitter iodine-131 only has a half-life of about 8 days and could not be measured directly anymore in studies conducted years later. A way to alleviate this problem is to make use of a so-called *retrospective dosimetry* technique, where the long-lived fission product iodine-129 with a half-life of almost 16 million years is measured instead. Based on the knowledge of several parameters, among them the ratio of the two iodine isotopes in the Chernobyl fallout and the results of pre-Chernobyl environmental measurements of iodine-129, one can derive thyroid doses due to iodine-131 from current iodine-129 activity measurements in the environment (Michel et al. 2005).

Environmental measurements of radioactivity have also been used in other epidemiological studies, possibly most prominently in the analysis of lung cancer associated with the radioactive noble gas radon (Darby et al. 2005). Here, the measured activity concentration (in becquerel per cubicmeter) of radon in the air of residential homes was used as a replacement for an actual dose quantity.

In the context of occupational radon exposure among miners, one often also encounters the term *working level month* (WLM) as a dose substitute. A working level (WL) is any combination of the radionuclide radon-222 and its decay products in 1 liter of air that emits a potential alpha radiation energy of $1.3 \cdot 10^5$ megaelectronvolt (MeV). This is equivalent to a radon-222 activity concentration in equilibrium with its decay products of $3.7\,\text{kBq m}^{-3}$. One WLM is then the radon exposure of a worker during 1 month of work (170 hours) at 1 WL.

One example of the extensive use of the dosimetric concept of the WLM has been the cohort study of German male uranium miners of the Soviet-German incorporated company Wismut. To assess the radiation exposure in this cohort study, a job-exposure matrix (JEM) was used. The use of JEMs is a common approach in occupational epidemiological studies (see chapters ▸Exposure Assessment and ▸Occupational Epidemiology of this handbook). The Wismut-JEM incorporates information from historical and concurrent measurements of radon gas and radon progeny as well as information on working conditions and mine architecture, to provide, in the end, exposure data for the years 1946–1989 for more than 900 different types of jobs, various mines, and workplaces (Grosche et al. 2006). Since the individually available information on the jobs being done, including times of absence, was mostly well documented, individual exposure could be estimated even for early periods.

In the context of the Life Span cohort Study (LSS), an individual dosimetry system for cohort members within 10 kilometers of the explosions was devised. Over the decades, a succession of dosimetry systems has been developed, the earliest dating back to the 1950s and consisting of just gamma and neutron "air doses" depending on the distance from the explosion and with simple corrections for shielding (Cullings et al. 2006). The recent DS02 dosimetry system, however, is much more sophisticated as will be described briefly below.

The goal of the dosimetry system was the reconstruction of the radiation dose consisting of gamma and neutron exposure for every cohort member. The whereabouts relative to the explosion were derived from personal questionnaire-based interviews. Certain model assumptions had to be made, including physical details of the different atomic bombs in Nagasaki and Hiroshima, iso-contours of the air density, as well as human phantoms. The organ doses were calculated by complex Monte Carlo simulations aiming to describe the transport of radiation from the explosion epicenter to the target organs.

The experimental verification of the methodology included Nagasaki-type test explosions at the Nevada test site with reconstructed Japanese houses to evaluate shielding conditions. These experiments were even continued after the test ban treaty in the early 1960s by installing neutron and cobalt-60 sources on a tower 465 m high (Rühm 2003). Furthermore, samples of building materials and minerals were taken in Nagasaki and Hiroshima to retrospectively evaluate the neutron and gamma radiation by means of the accelerator mass spectrometry detection of neutron activation products or beta or gamma spectrometry and the thermoluminescence technique, respectively. Such extensive efforts are not possible in all radiation epidemiology studies, but experiences obtained in the Life Span Study have been important for many subsequent investigations.

Another important range of studies pertains to the Mayak nuclear facility located near Chelyabinsk in Russia. This facility began operations in 1948 to produce plutonium for the Soviet nuclear weapons program. Substantial exposure to external radiation and to internally deposited plutonium for both workers and the general public occurred mainly in the 1940s and 1950s. Three cohorts were formed: the cohort of Mayak nuclear workers, which were exposed externally at work to neutron and gamma radiation as well as internally due to plutonium inhalation; the cohort of Techa river residents, who were exposed internally due to ingestion of cesium-137 and strontium-90 as well as externally due to deposited radioactive material; the Techa river offspring cohort, with an exposure pattern similar to that of the Techa river cohort. As in the LSS, the retrospective dose assessment in the Mayak study consisted of several complementary approaches. External exposure could partly be monitored with individual film badge dosimeters (Vasilenko et al. 2007). For all cohorts, electron spin resonance spectrometry and strontium-90 beta spectrometry measurements of teeth enamel were conducted to determine absorbed doses. Furthermore, whole-body counting measurements were conducted for all cohorts. Similar to the LSS, thermoluminescence measurements of ceramic building materials helped to estimate external exposures. Biokinetic modeling approaches, partly in combination with urine monitoring data, were employed to cover strontium-90 and plutonium metabolism effects, that is, the distribution and retention of these radionuclides in different parts of the body, as well as to provide lung dosimetry.

Biodosimetry involves the use of biomarkers to assess exposure to ionizing radiation. The radiobiological investigation of cellular effects, in particular induced chromosomal aberrations, has contributed substantially to the understanding of radiation as a toxic agent and cancerogen (National Research Council 2006). Unstable aberrations such as dicentric and ring chromosomes are useful to assess rather recent exposures, whereas stable aberrations including insertions and translocations can provide information about exposures over longer time periods. However, the validity of stable chromosomal aberrations as a retrospective biomarker of radiation exposure is still under debate and investigation. The lower limit of sensitivity of most available techniques is currently above typical low level exposures of many populations under study. New molecular biomarkers have been developed in recent years, but their validity and usefulness for radiation epidemiology is not yet established (Pinto et al. 2010).

In summary, many different ways of measuring, modeling, or estimating radiation doses in epidemiological studies are employed depending on the specific circumstances of the respective studies. Compared to the common indirect exposure assessment in epidemiological studies, it is an outstanding feature of studies with direct measurements of radiation doses that exposure can be assessed with a comparatively high accuracy. Nevertheless, uncertainty in exposure assessment remains an important problem in radiation epidemiology. Improved exposure assessment combined with statistical approaches will help to further quantify and reduce dosimetric and other uncertainties in the future.

50.4 Analysis of Radiation Epidemiology Data

Biological and epidemiological aspects of carcinogenesis of ionizing radiation have been studied extensively for more than 100 years. Today, the main aim of the statistical analysis of new epidemiological studies is not to provide estimates of cancer risk in exposed versus those of non-exposed persons but to quantify the risk as a function of dose and dose rate of high-LET (linear energy transfer) and low-LET radiation and to investigate modifying effects on the dose-response curve such as age at exposure, sex, ethnic group, and lifestyle factors. It is also known that the relationship between risk (or incidence) and dose varies for different types of cancer, and distinctions between different types of radiation are necessary (see Sect. 50.3.1). It has been postulated, for example, by Upton (1977) that the process of carcinogenesis is a combination of initiation (at low to medium dose, e.g., by a mutation) and the survival of the cells; at high doses, the cell is killed by radiation and cannot develop into cancer. As individual doses are available or estimates can be provided by adequate dosimetry in many exposure situations, complex modeling has a long tradition in radiation epidemiology. For X-rays, gamma rays, and electrons, the dose-incidence curve $I(D)$ can in general be described as follows:

$$I(D) = (\lambda_0 + \alpha D + \beta D^2)\exp(-\gamma D + \delta D^2),$$

where λ_0 is the incidence of non-exposed persons (which may depend on further factors such as age and sex). Here, α, β, γ, δ are constant coefficients and D is the dose. More sophisticated expressions include dependency of the parameters $(\alpha - \delta)$ on modifying factors. It has been noted that the radiation-induced excess may be approximated by a constant percentage of the natural age-specific incidence at the time of radiation exposure (relative risk model) or by a constant number of additional cases (absolute risk model).

While in other fields of epidemiology the relative risk (*RR*) is used more frequently to quantify risk, the additional effect of radiation has traditionally been in the focus of analyses in radiation epidemiology. This led to the use of "excess absolute risks" (*EAR*) and "excess relative risks" (*ERR*) to better communicate the risk associated with radiation exposure of populations.

The *EAR* is the difference between the absolute risk of exposed persons and the absolute risk among unexposed persons that is

$$EAR(d, a, e) = \lambda(d, a, e) - \lambda_0(a, e),$$

where $\lambda(d, a, e)$ is the risk (or incidence) of persons with dose $= d$, attained age $= a$, and age at exposure $= e$ and $\lambda_0(a, e)$ is the corresponding baseline rate ($d = 0$).

The excess relative risk (*ERR*) is just the relative risk -1 or

$$ERR(d, a, e) = [\lambda(d, a, e)/\lambda_0(a, e)] - 1.$$

Of course, more complex forms for the risk $\lambda(d,a,e,x)$ can be used, and the risk can be treated as a more complex function of dose (e.g., including a quadratic term) or including further covariates (x). Standard mathematical methods can be used to estimate the parameters in the model and were already used to analyze the data from the Japanese atomic bomb survivors (see Sect. 50.5.4) prior to the extensive use of relative risk models in other epidemiological areas.

In addition to their application for the analyses of data from the atomic bomb survivors, models as described above have been used in studies of the nuclear industry workers, medically exposed persons, and many other cohort and case-control studies in the past. More recently, a new family of models – so-called mechanistic models (Jacob and Jacob 2004; Eidemüller et al. 2010) – was introduced, and several large data sets were reanalyzed with these models. Results from the traditional (now called empirical models) and the mechanistic models are now combined to estimate individual risks to develop cancer and will be used for radiation protection purposes. Mechanistic models include separate parameters for the initiation and promotion steps in cancer development. However, as many processes leading to the different types of cancer are not well understood so far, these models can only include selected aspects of the carcinogenic process.

A major issue in the analysis of radiation epidemiology data is the estimation of risks in the low dose range, that is, below 100 or 200 milligray (mGy). This dose range is of particular interest since most occupational and population exposures occur at low and protracted doses, that is, doses spread over long periods of time. This range is also of great importance for radiation protection.

The shape of the dose-effect curve for solid cancer at lower doses is assumed to be linear without a threshold dose, the so-called linear non-threshold (LNT) hypothesis. This hypothesis includes linear extrapolation from effects of moderate to high doses of ionizing radiation down into the low level range. In this dose range, direct effect estimates often are less precise, such that the actual shape of the curve is frequently driven by estimates at higher doses (see Fig. 50.2, Sect. 50.5.4).

In addition, biological and other considerations suggest that the curve could well take another form in the low dose range; however, statistical testing indicates that other, for example, non-linear or threshold models do not describe the data for solid cancer significantly better than the linear model (Preston et al. 2007). As the nature of the exposure and many other issues related to the study of atomic bomb survivors are unique, direct assessments of radiation risks from low and protracted exposures have been performed in many different study populations in order to arrive at an empirically well-founded risk assessment for ionizing radiation at low and protracted doses.

When comparing data across studies, transferring risk estimates from one population to another is a major issue. This has been of particular concern with regard to the data of the atomic bomb survivors. For example, the Japanese population is known to have markedly higher stomach cancer risks than US Americans; conversely, US American women have much higher breast cancer risks than Asian women. Different baseline rates of disease in populations obviously influence the number of excess cases induced by radiation. The validity of transferring the risk estimated in one population to estimate the disease burden associated with radiation exposure in a different population remains an issue of further investigations (UNSCEAR 2000).

50.5 Selected Radiation Epidemiology Studies

The effects of ionizing radiation have been studied in many different exposure situations, each posing specific challenges to the planning, conduct, analysis, and interpretation of the study. This section presents selected epidemiological studies as examples of the various exposure situations and the populations studied in radiation epidemiology.

50.5.1 Occupational Radiation Studies

There is a long history of radiation studies in occupational settings. Among the early radiation epidemiology studies, the investigation of radium-dial painters has received widespread attention. Women employed in the US radium-dial industry before 1930 used their lips to make fine points on their paint brushes. This habit resulted in the ingestion of large amounts of radium-226. Radium was deposited in the bones with the consequence of excess bone sarcomas many years later (Stebbings et al. 1984).

Many other occupational studies, mainly with a focus on cancer have been conducted since. Some broad categories of occupational groups are presented in Table 50.4, and respective studies will be discussed below:

50.5.1.1 Airline Crew

Airline crew (and others traveling in aircraft) are exposed to increased levels of cosmic radiation. Commercial pilots and flight crews flying about 1,000 h per

Table 50.4 Overview of occupational groups and their exposure situation

Occupational group	Main exposure	Specification
Airline crew	Cosmic radiation	External exposure
Uranium miners	Radon (^{222}Rn)	Internal exposures
Medical staff	X-rays	External exposure
Nuclear industry workers	Gamma radiation	External exposure
Workers in nuclear weapon industries (e.g., Mayak)	Gamma radiation Plutonium (^{239}Pu)	External exposure Internal exposures
Chernobyl accident workers	Gamma radiation Iodine (^{131}I)	External exposure Internal exposure

year receive annual doses of 1–6 mSv, depending on flight routes and several other factors. Cosmic radiation includes a substantial neutron component, and thus aircrew is one of the few populations exposed to neutron radiation. Since about 1980, numerous mortality and incidence studies among aircrew have been conducted. One of the largest is the pooled European cohort study of airline crew with data from nine European countries. Radiation exposure was crudely estimated based on flight routes and other occupational data. Low cancer and cardiovascular mortality was found, but there also were several cancer sites with risk elevations, such as malignant melanoma of the skin among pilots and male cabin crew and, although less prominently, breast cancer mortality among female cabin crew (Standardized Mortality Ratio, *SMR*: 1.11, 95% confidence interval: 0.82–1.48) (Blettner et al. 2003; Zeeb et al. 2003). Studies of cancer incidence yielded slightly higher breast cancer risk elevations (Sigurdson and Ron 2004). Possible selection effects and problems in confounder control, among others, complicate the evaluation of radiation effects in this occupational group, but the overall impact of cosmic ionizing radiation on cancer incidence and mortality appears to be small (Hammer et al. 2009a).

50.5.1.2 Underground Hard-Rock Miners

Underground hard-rock miners such as uranium, iron, tin, and gold miners are exposed to alpha particles emitted from radon decay products. This results in considerable doses to lung tissue when the particles attach to dust and remain in the lung (Boice and Lubin 1997). A pooled analysis of 11 cohorts of underground miners in which radon exposure was estimated for individual workers included 65,000 men accruing nearly 1.2 million person years of observation and over 2,700 deaths from lung cancer. The excess relative risk of lung cancer was approximately directly proportional to the cumulative exposure to radon, that is, there was a linear exposure-disease relationship (Lubin et al. 1995). A major issue in radon and lung cancer studies is the fact that tobacco smoke is the leading cause of lung cancer. Hence, it is important to determine the overall risk arising from the combined exposure to tobacco smoke and radon. In the pooled miner analysis, the excess relative risk per WLM (*ERR* WLM^{-1}) coefficient for never-smoking miners was three times the *ERR* WLM^{-1} coefficient for ever-smoking miners. About 40% of all lung cancer deaths were attributed to radon-progeny exposure, 70% in never-smokers, and 40% in smokers (Lubin et al. 1995; Wakeford 2009).

Extensive uranium extraction took place from 1946 until 1990 at the Wismut mining company in the former German Democratic Republic. About 59,000 male former employees of this company form the largest single uranium miner cohort worldwide. The cohort has been followed up for mortality since the beginning of 1946 to the end of 2003 with a total of 3,016 lung cancer deaths and two million person years. There was a twofold increase in the mortality for lung cancer, and the *ERR* WLM^{-1} coefficient for lung cancer was estimated as 0.0019 (95% confidence interval (CI): 0.0016–0.0022) (Walsh et al. 2010). This is somewhat lower than the earlier estimates from the pooled study mentioned above.

50.5.1.3 Medical Workers

Within a few years after the discovery of X-rays in 1895, there were reports of dermatitis, ulceration, skin cancer, sarcomas, other serious radiation-induced injuries, and deaths in radiologists (Frieben 1902; Hesse 1911). An association between employment as a radiologist and increased risk of acute myeloid leukemia was confirmed in a 1944 report (March 1944), a year before the atomic bombings in Japan.

The results of 100 years of observation on British radiologists (Berrington et al. 2001) showed an excess of cancer mortality of about 40% for radiologists who had been working for more than 40 years compared with other medical practitioners. This was considered as a probable long-term effect of radiation exposure in those who first registered before 1954. There was no evidence of an increase in cancer mortality among radiologists who first registered after 1954. Most likely, the occupational radiation exposures of those employed in more recent years have been lower due to improved radiation protection.

Another example of a medical worker study is the investigation of cancer incidence and mortality among over 90,000 US radiological technologists (77% of whom were women) who were certified for at least 2 years between 1926 and 1982. Regarding the incidence of cancer diseases, female technologists had an elevated risk for all solid tumors including breast cancer. Both female and male technologists showed elevated risks for melanoma and thyroid cancer (Sigurdson et al. 2003), and among workers employed in early years, there was evidence for increased leukemia mortality (Linet et al. 2005).

50.5.1.4 Nuclear Industry Workers

Nuclear industry workers form a very suitable study population to assess effects of low and protracted radiation exposure, as worldwide several hundred thousand employees with detailed personnel and radiation dose records are available for studies (Wakeford 2005, 2009). The *United Nations Scientific Committee on the Effects of Atomic Radiation* (UNSCEAR 2010) defines low doses as those of 100 mGy or less and low dose rates as 0.1 mGy per minute or less (averaged over 1 h) for external X-rays and gamma radiation. The predicted low dose excess radiation-related risk of cancer is small. On the other hand, low and protracted doses are relevant for most persons exposed in an occupational context as well as for the overall population.

The International Agency for Research on Cancer (IARC) coordinated the largest nuclear industry worker study to date (Cardis et al. 2005, 2007). This multinational retrospective cohort study conducted in 15 countries included approximately 400,000 predominantly male workers who had been employed for at least 1 year and who were monitored for external photon radiation using personal dosimeters. The mean cumulative individual dose was 19.4 mSv with 90% of workers having a dose lower than 50 mSv.

For mortality from all cancers excluding leukemia, the excess relative risk was 0.97 Sv^{-1} and was significantly different from zero (95% CI: 0.14–1.97). The *ERR* for mortality from lung cancer was significantly increased with 1.86 Sv^{-1} (95% CI: 0.26–4.01). Since the observed risk of mortality from all cancers excluding leukemia was twice as high as the risk estimate for solid cancer mortality observed in the LSS, the results were intensively discussed.

This study is a revealing example for the application of sensitivity analyses in epidemiological studies which can help to identify, albeit post hoc, specific data constellations, design issues, or, at times, systematic errors. In the 15-country study, the Canadian data have a surprisingly large influence on the *ERR* coefficient for all cancers other than leukemia, even though there are just over 200 Canadian cancer deaths (4% of the total number of cancer deaths). Exclusion of the Canadian cohort from the study reduced the estimate by 40% and changed it from statistically significant to statistically non-significant. Among the Canadian cohort, a specific subgroup of 3,088 workers employed before 1965 was the only group of workers with a radiation-associated increase in risk of solid cancer mortality, and this group had a strong impact on the study findings. A reanalysis showed that these workers had incomplete dose information. A potentially significant gap in reporting zero doses was discovered and introduced substantial distortion of the dose-effect relationships (Ashmore et al. 2010).

Concerning leukemia (excluding chronic lymphocytic (CLL) leukemia which is often considered not to be induced by radiation), the *ERR* per dose unit was 1.93 Sv^{-1} (95% CI: <0–8.47), indicative of an excess radiation-related risk of mortality from leukemia. Given observations from the Mayak Worker Studies (see next section), where the increased leukemia risk was concentrated in the period 3–5 years after the dose was received (Wakeford 2009; Shilnikova et al. 2003), temporal patterns of leukemia risk might be different after low dose protracted exposures compared to higher cumulative doses.

In some countries, worker registries provide an excellent basis for epidemiological studies. For example, in the latest follow-up of 174,541 persons from the UK National Registry for Radiation Workers, an increased risk for leukemia mortality and incidence was observed. Consistent with other studies, the leukemia subtype showing the strongest evidence of an association with radiation is chronic myeloid leukemia (Muirhead et al. 2009). Nuclear workers continue to be studied, and results from additional cohorts will be available in the future.

50.5.1.5 The Mayak Nuclear Weapon Worker Studies

In addition to exposure to external sources of ionizing radiation, some nuclear industry workers inhale or ingest radioactive material, leading to a more complex exposure situation that also affects comparability with other groups. In particular, workers in nuclear weapon industries are exposed to uranium, plutonium, and tritium. The example of Mayak workers has been described in Sect. 50.3. External radiation exposures of these workers far exceeded the doses of other nuclear worker cohorts. Among workers hired before 1959, the mean cumulative external dose was 1.2 Gy, which is more than an order of magnitude above the mean exposure in the 15-country study (see Box 50.1).

There were also substantial internal exposures for plutonium. The mean lung dose was 0.21 Gy, the mean liver dose was 0.28 Gy, and the mean bone surface dose was 1.0 Gy (Vasilenko et al. 2007).

Sokolnikov et al. (2008) published findings for solid cancers from the cohort of 17,740 Mayak workers first employed during 1948–1972. A total of 13,816 workers were potentially exposed to plutonium. Some of the main results are shown in Table 50.5.

The higher ERR coefficients for females, notably for lung cancer and liver cancers, might be due to higher background risks in males as a result of higher tobacco and alcohol consumption. The increased leukemia risk was concentrated in the period 3–5 years after the exposure (Shilnikova et al. 2003). So far, there was no indication of any significant effect of plutonium exposure upon leukemia mortality.

Other examples of occupational groups exposed to plutonium are workers in the Manhattan Project (Los Alamos, USA, 1944–1945; the Manhattan Project was a program that produced the first atomic bomb during the World War II and was initiated by the USA with participation from the UK and Canada), workers at the Rocky Flats nuclear weapon production facility (Colorado, USA, 1952–1979), the Hanford nuclear facility with the world's first plutonium production reactor (Washington, USA, 1944–1978), or the Sellafield plutonium workers (UK, 1947–1975).

50.5.1.6 Chernobyl Cleanup Workers

After the accident at the Chernobyl nuclear power plant in April 1986, so-called liquidators participated in the cleanup of the reactor, sarcophagus construction, decontamination, and other recovery operations. About 240,000 persons worked at

Box 50.1. Comparison of mean cumulative external dose in two important occupational studies

Nuclear energy workers in the 15-country study: 19.4 mGy
Mayak workers hired before 1959: 1,200 mGy

Table 50.5 Risk estimates for different cancer sites derived from the cohort study of Mayak nuclear facility workers (Sokolnikov et al. 2003, 2008)

Cancer	*ERR* coefficients for external dose (adjusted for internal dose)	*ERR* coefficients for tissue-specific doses for the internal exposure to plutonium, adjusted for external dose
Lung cancer	0.19 Gy^{-1}(95% CI: 0.05–0.39)	Males: 7.1 Gy^{-1} (95% CI: 4.9–10) Females: 15.0 Gy^{-1} (95% CI: 7.6–29)
Liver cancer	0.21 Gy^{-1} (95% CI: <0–1.0)	Males: 2.6 Gy^{-1} (95% CI: 0.7–6.0) Females: 29.0 Gy^{-1} (95% CI: 9.8–95.0)
Bone cancer	0.35 Gy^{-1} (95% CI: <0–4.4)	Males: 0.76 Gy^{-1} (95% CI: <0–5.2) Females: 3.4 Gy^{-1} (95% CI: 0.4–20.0)
Leukemia (excl. CLL)	3–5 years since dose received: 6.9 Gy^{-1} (90% CI: 2.9–15) >5 years since dose received: 0.5 Gy^{-1} (90% CI: 0.1–1.1)	

the site during 1986–1987 when exposures were the highest. Sent to the reactor site usually for a period of about 2 weeks, these cleanup workers were exposed primarily to external gamma radiation. Estimates from the national Chernobyl registry data in Ukraine, Belarus, and the Russian Federation indicated a mean dose from external radiation of 144 mGy in 1986, 90 mGy in 1987, and 36 mGy in 1988–1989. The mean effective dose for cleanup workers is estimated to have been around 100 mSv. Furthermore, Chernobyl workers received substantial thyroid doses from intake of radioactive iodine (Wakeford 2009; Romanenko et al. 2008).

In the USSR, a comprehensive registration and active follow-up system was set up for the persons most affected by the Chernobyl accident. However, incompleteness of data, the lack of verification and quality control, and selection bias pose problems. In Russia and the Ukraine, no centralized cancer registration system was in place at the time of the accident, whereas in Belarus a computerized National Cancer Registry existed since 1970. The dosimetric information available for cleanup workers is derived from personal dosimeters, group dosimetry, and measurements at various points where liquidators worked (National Research Council 2006).

Ivanov et al. (2008) conducted a cohort study of 103,427 Chernobyl emergency workers with a follow-up from 1986 to 2003. A total of 22,408 persons were engaged in the early recovery operations in April–July 1986. They received the highest external doses (mean dose: 168 mGy), while the mean dose for those emergency workers employed in 1988–1990 was 33 mGy. A total of 87 cases of thyroid cancers were observed with a statistically significant overall *SIR* of 3.47 (95% CI: 2.80–4.25). The highest *SIR* was found for the emergency workers involved in the recovery operations from April to July 1986. In this group, 34 cases of thyroid cancers occurred, and the *SIR* value was 6.62 (95% CI: 4.63–9.09). Further work is ongoing, and epidemiological data from Chernobyl related studies will continue to enhance the knowledge about radiation-associated disease risks for many years to come.

50.5.2 Environmental Radiation Studies

Not all epidemiological study designs are equally informative for the estimation of radiation risk in humans. Ecological studies do neither include individual exposure estimates nor individual confounder information. In ecological studies, aggregated population estimates or distance-based exposure proxies are used in order to define the likelihood for exposure. Thoroughly conducted case-control or cohort studies overcome these disadvantages. For this reason, this section focuses on selected examples of current case-control or cohort studies that exemplify research approaches in given situations of public health relevance. Comprehensive overviews of epidemiological studies on environmental radiation are given, for example, by the National Academy of Sciences Advisory Committee on the Biological Effects of Ionizing Radiation (National Research Council 2006) and by UNSCEAR in its regular reports.

50.5.2.1 Populations Exposed to Fallout from the Chernobyl Accident

The explosion at the Chernobyl plant in 1986 released substantial quantities of radionuclides into the atmosphere, resulting in the contamination of a large geographical area. Initially, exposures of the population were principally due to radioisotopes of iodine-131 and subsequently cesium-137 from both external exposure and the consumption of contaminated milk and other foods.

Numerous studies have shown an association between radiation exposure from Chernobyl and an increase in the incidence of thyroid cancer in Belarus, Ukraine, and Russia (National Research Council 2006). Iodine deficiency seems to be an important modifier of the risk of radiation-induced thyroid cancer. Some regions contaminated by the Chernobyl accident are areas of mild to moderate iodine deficiency which enhances the risk of thyroid cancer following irradiation. Furthermore, exposure at younger ages is associated with the greatest risk for thyroid cancer. Three case-control studies on thyroid cancer in children and young adults exposed to the Chernobyl accident have shown that exposure at younger ages is associated with the greatest risk for thyroid cancer (UNSCEAR 2008), while rates of childhood leukemia seemed unchanged after Chernobyl.

50.5.2.2 Techa River Populations

The exposure of persons living near the Techa River in Russia consisted of a mixture of radionuclides, mainly strontium-90 and cesium-137. Approximately 30,000 persons living near the Techa River, among them many children and adolescents at the time of exposure, have been followed in a cohort for about 5 decades. Individualized stomach and red bone marrow doses have been estimated for each cohort member. For example, the average stomach dose for cohort members is estimated to be around 0.03–0.04 Gy, with maximum organ doses of about 0.5 Gy. Significantly elevated risks of leukemia and solid cancer mortality have been demonstrated in this study (Krestinina et al. 2007; Eidemüller et al. 2010).

50.5.2.3 Leukemia in Children Living in the Vicinity of Nuclear Power Plants

Leukemia is the most common malignancy in childhood (Parkin et al. 1988). The causes of most leukemias are still unknown. However, high doses of ionizing radiation are an established risk factor for childhood leukemia (Lightfoot and Roman 2004). Regarding low and protracted doses, there has been a long-standing debate whether or not the emission of ionizing radiation during routine operation of nuclear plants increases the risk of leukemia in children living near the plants.

Numerous studies on this issue have been conducted in Europe and the USA, usually employing ecological designs where cancer rates are compared between regions of decreasing vicinity to nuclear power plant sites. A population-based case-control study was conducted in Germany where cases were children younger than 5 years (diseased between 1980 and 2003) registered with leukemia at the German Childhood Cancer Registry (GCCR) (Kaatsch et al. 2008). Population-based controls were matched (1:3) for date of birth, age at diagnosis, sex, and nuclear power plant area (at the date of diagnosis) of the corresponding case. Residential proximity to the nearest nuclear power plant was determined for each subject individually. The study included 593 leukemia cases and 1,766 matched controls. The main result was a negative trend for distance, indicating that cases lived closer to nuclear power plants than the randomly selected controls. A categorical analysis showed a statistically significant odds ratio of 2.19 (lower 95% confidence limit (CL): 1.51) for residential proximity <5 km compared to residence outside this area. A French case-control study of acute leukemia reported similar findings (Sermage-Faure et al. 2012), whereas in a nationwide Swiss cohort study, no excess leukemia risk among children living near nuclear power plants was seen (Spycher et al. 2011). As the radiation exposure near a nuclear power plant in routine operation is very small compared to exposure of the general public from other sources (e.g., terrestrial or medical radiation), other causes beyond radiation exposure have to be considered.

50.5.3 Medical Radiation Studies

Diagnostic radiation techniques include X-ray radiography, fluoroscopies, CT, or interventional radiology. As indicated earlier, medical radiation increasingly contributes to the overall population radiation exposure. Advances in radiation protection help to lower the doses per individual examination; however, since radiological imaging is continuously becoming more widely available and informative with regard to the detection of abnormalities, the balance of benefits and risks requires continuous evaluation. Furthermore, the radiation exposure of patients could be part of their treatment. Usually, therapeutically exposed persons are not typical for the general population, for example, with respect to age (e.g., cancer patients). In addition, exposed individuals have medical conditions that influence the trade-off between the benefits and the risks of using radiation. Nevertheless, the detailed study of potential adverse effects of medical usage of radiation is essential to further gain insight on the benefit-risk balance and to assure appropriate radiation protection of patients. Such studies also provide opportunities to understand radiation exposure effects in specific groups and ages. Some case-control and cohort studies highlighting the respective epidemiological approaches are presented below.

50.5.3.1 The British Childhood Cancer Survivors

Children and young adults treated for cancer nowadays have a very good survival prospect but may face late effects of their treatment. The British Childhood Cancer Survivor Study is a population-based cohort of 17,981 individuals diagnosed with childhood cancer (1940–1991) and surviving for at least 5 years, followed up through December 2006. The greatest excess risk associated with subsequent primary neoplasms at age above 40 years was found for patients with primary digestive and genitourinary neoplasms (Hawkins et al. 1987; Reulen et al. 2011). A joint analysis of the British and French cohorts (de Vathaire et al. 1989) indicated that the risk for developing a second cancer in the 25 years after diagnosis of the first cancer was 12% (Tucker et al. 1991).

Numerous studies of second cancer following radiotherapy in adults have focused on patients with a favorable long-term prognosis (e.g., cervical cancer, breast cancer, testicular cancer, thyroid cancer, or Hodgkin's lymphoma). It should be noted that chemotherapy or hormonal therapy used in the treatment of cancer patients is an important confounder in investigations of the relationship between radiation therapy and second primary cancer. Concerning benign diseases, several cohorts of children and adults treated with ionizing radiation have been followed up over long periods (Kleinerman 2006).

50.5.3.2 The Oxford Survey of Childhood Cancers

Exposures to children including intrauterine exposures during pregnancy are of particular concern in this respect. An association between childhood leukemia and an abdominal X-ray examination of the pregnant mother was first reported in the 1950s (Steward et al. 1956, 1958). It was obtained from a large case-control study conducted in the United Kingdom in which the mothers of children who died from cancer and mothers of control children were asked about their child's history of radiographic examinations in utero and after birth. In 1981, the Oxford Survey of Childhood Cancers (OSCC) included 15,276 case-control pairs. For the calendar period 1953–1981, the overall excess risk of childhood cancer associated with maternal exposure to abdominal radiography during pregnancy in the Oxford survey was approximately 40% (odds ratio (OR) = 1.39, 95% CI: 1.30–1.49) (Gilman et al. 1989).

In contrast to these findings, no statistically increased leukemia risks were seen among children exposed in utero in Hiroshima and Nagasaki, but the numbers were very small. There are numerous arguments against and in favor of the validity of the OSCC findings (Doll and Wakeford 1997; National Research Council 2006). Based on an evaluation of all available evidence, Wakeford and Little (2003) concluded that doses to the fetus in utero of the order of 10 mGy can already lead to increases in the risk of childhood cancer.

50.5.3.3 Other Studies on Diagnostic Radiology

Postnatal medical diagnostic exposures to children have also been studied extensively, albeit with inconsistent results. Recently, the RICC study (Radiation-Induced Childhood Cancer Study) focused on the postnatal exposure of children (<15 years) to X-ray examinations and the risk of childhood cancer. The cohort included 92,957 children who had been examined with diagnostic X-rays in a large German hospital during 1976–2003, and individual dose estimations for all radiological examinations performed in this hospital were available. Newly diagnosed cancers occurring between 1980 and 2006 were determined through record linkage to the German Childhood Cancer Registry. The median radiation dose was very low (0.007 mSv). Eighty-seven incident cases were found in the cohort. The SIR for all cancers was 0.99 (95% CI: 0.79–1.22). No trend in the incidence of total cancer, leukemia, or solid tumors with increasing radiation dose was observed in the SIR analysis or in the multivariate Poisson regression. This result was in contrast to several earlier studies, but the RICC risk estimates had wide confidence intervals and were compatible with a broad range of risks (Hammer et al. 2009b).

During the past 20 years, a rapid increase of pediatric computed tomography (CT) use could be observed in many countries. The CT radiation doses are relatively high compared to those from most conventional X-ray examinations. An international cohort study in 11 European countries, coordinated by the IARC, studies the health consequences of ionizing radiation exposure from CT in childhood and adolescence. This *EPI-CT Study* will eventually include about one million children. The enrolment period began in 1984 with a follow-up until about 2010. Cohort data from the UK were published in 2012 and indicate a risk increase for childhood leukemia associated with ionizing radiation from CT comparable to that estimated from the LSS data. For brain tumors, a markedly larger risk increase was found in this study (Pearce et al. 2012).

Most diagnostic medical radiation is, however, applied to adult patients. An example for a study of relatively high diagnostic exposures is the investigation of patients monitored during their pneumothorax treatment with repeated fluoroscopic examinations (see Table 50.3) (Howe and McLaughlin 1996). In modern medicine, CT examinations as well as interventional radiology procedures are among the interventions leading to relatively high organ doses. The attributable lifetime cancer mortality for a 45-year-old adult undergoing multiple full-body CT scans has been estimated to be around 2% (Brenner and Elliston 2004). Similarly, the use of radionuclides in nuclear medicine – both for diagnostic and therapeutic purposes – has been investigated with regard to health risks. These studies generally point to small risk increases. For example, in the UK, annually less than 20 solid cancers and 12 cases of leukemia are estimated to be due to nuclear medicine procedures, as against a total of about 313,000 cancers diagnosed annually in the country (Parkin and Darby 2011).

50.5.4 Atomic Bomb Survivor Studies

The Life Span Study (LSS) of survivors of the atomic bombings in Hiroshima and Nagasaki as the key study of radiation epidemiology has been mentioned in several earlier sections. The LSS is a representative cohort of all ages and both sexes, with little selection bias concerning socioeconomic factors, diseases or occupations, and a follow-up since 1950. The LSS comprises a wide range of radiation exposures and a relatively accurate dosimetry that has undergone constant improvements over time. The exposure is a whole-body exposure, which allows the assessment of risks for specific cancer sites (National Research Council 2006). What has to be noted, however, is the fact that the studies of atomic bomb survivors only provide evidence for risks following a single acute exposure. Whether acute and protracted exposures lead to similar effects in humans is not well understood, and there is an ongoing controversy whether doses from acute and chronic exposures should be treated differently (Jacob et al. 2009).

The LSS cohort is comprised of 120,321 people who were identified at the time of the 1950 census. Cohort members include three major groups of registered Hiroshima and Nagasaki residents: (1) about 54,000 atomic bomb survivors who stayed within 2.5 km of the hypocenter at the time of the bombings; (2) about 40,000 atomic bomb survivors who stayed between 2.5 and 10 km of the hypocenter at the time of the bombings; and (3) about 27,000 people who resided in the cities at the time of the census but did not stay there at the time of the bombings (Preston et al. 2007). Vital status is updated in 3-year cycles through the Japanese family registration system. Death certificates provide data on the cause of death. Since 1990, data from Hiroshima and Nagasaki tumor registries were linked to the LSS cohort, which allows the evaluation of cancer incidence. The Adult Health Study (AHS) is a cohort which includes a 20% subsample of the LSS. AHS subjects were invited to participate in biennial comprehensive health examinations. The AHS thus complements the LSS and provides more detailed health endpoints based on clinical data (National Research Council 2006).

The LSS provides risk estimates for solid cancers as well as for hematological cancers. Mortality and incidence data are available from reliable registration sources, and due to the size and duration of follow-up of the cohort and the variability of doses, the precision of risk estimates from the LSS is relatively high for many outcomes.

In terms of solid cancer incidence and mortality, risk increases have been found for numerous individual cancer types and for solid cancer overall, and the dose-response curve generally follows a linear pattern (see Fig. 50.2). Latency periods of 10 years or more have been observed for most solid cancers, and the highest risks were seen for those exposed when they were young. The most recent overall risk estimates indicate an incidence increase of 0.35 Gy^{-1} (90% CI: 0.28–0.43) for men and 0.58 Gy^{-1}(90% CI: 0.43–0.69) for women, assuming an age of 30 years at exposure and an attained age of 70 years (Preston et al. 2007). The excess mortality risk is in a very similar order (0.47 Gy^{-1} for both sexes combined) (Preston et al. 2003). For leukemia mortality, an average excess relative risk of 4.04 at 1 Sv has been found in the LSS (Preston et al. 2004), and the proportion of leukemia deaths in the cohort attributable to the radiation (around 50%) is much larger than the comparable proportion for solid cancers.

The LSS is also an important source of information on the radiation-related incidence or mortality risk from diseases other than cancer. Cardiovascular diseases

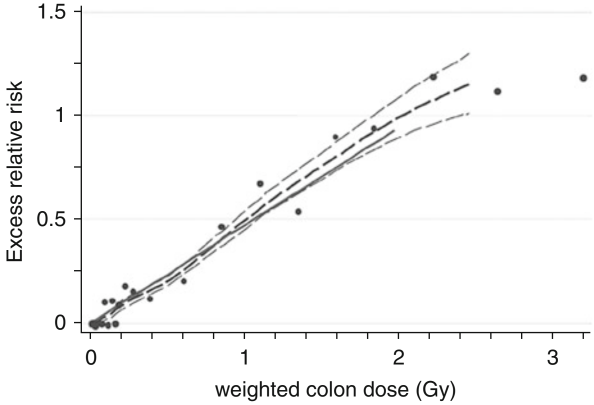


Fig. 50.2 LSS solid cancer incidence, excess relative risk by radiation dose, 1958–1998. The *thick solid line* is the fitted linear sex-averaged excess relative risk (ERR) dose response at age 70 after exposure at age 30. The *thick dashed line* is a non-parametric smoothed estimate of the dose category specific risks, and the *thin dashed lines* are one standard error above and below this smoothed estimate (From Preston et al. 2007 with permission)

have been consistently associated with ionizing radiation, although risks per dose seem to be about 1/4 of the comparable cancer risks. In the LSS, a 0.14 Gy^{-1} increase for heart disease mortality (95% CI: 0.06–0.23) and a somewhat lower risk increase of 0.09 Gy^{-1} for mortality from cerebrovascular disease were estimated (Shimizu et al. 2010). A number of other cohort studies indicate risk increases in a similar order. What remains somewhat unclear is the question of risk increases below doses of about 500 mGy, since the available results are inconsistent, and there are many confounding factors rendering the valid assessment of this question difficult. This is especially so when good confounder information, for example, on smoking or other cardiovascular risk factors possibly associated with radiation exposures, is missing, which is the case in the majority of studies with information on cardiovascular risks.

Beyond cardiovascular diseases, digestive and respiratory diseases (Preston et al. 2003) and eye lens cataract (Ainsbury et al. 2009) have been found to be associated with radiation exposure. Presently, there is substantial research interest in these non-cancer outcomes. For the development of cataract, for example, the previously assumed threshold dose of about 2 Gy is now considered too high, and there may be no threshold dose at all.

50.6 Radiation Epidemiology and Radiation Protection

Soon after the discovery of X-rays in 1895, it became evident that there was an urgent need to implement radiation protection rules to avoid adverse effects for exposed persons. As early as 1896, the importance of distance to the radiation source, the influence of duration of exposure, and the consideration for skin protection with Vaseline for the most exposed areas were discussed. Since then, radiation protection was continuously improved by technical means and by developing and updating protection standards. Changes were driven by new data on the biological effects of ionizing radiation, by large epidemiological studies, as well as by public concern (ICRP, International Commission on Radiological Protection 2009).

At the conferences of the American Roentgen Ray Society in 1924 and the International Congress of Radiology (ICR) in 1925, radiation hazard and protection was a major issue. First formal standards for protecting humans were proposed by the US Advisory Committee on X-Ray and Radium Protection in 1934. The limits were proposed mainly following observations of serious health effects in radium-dial painters.

At the second meeting of the ICR in 1928 in Sweden, the International Commission on Radiological Protection (ICRP, at this time IXRPC) was founded as a registered charity and as an independent non-governmental organization. Since then, the ICRP has played a central role in radiation protection, including evaluation of scientific studies, proposing methods of measuring ionizing radiation and units to compare different types of radiation, proposing dose limits, and discussing concepts and philosophy of protection.

In 1949, the concepts of “absorbed dose” and “dose equivalent” were introduced, and from 1950 onward, results from the LSS became available as well as extensive data from animal experiments. Based on these scientific foundations, the ICRP continued to recommend dose limits for a wide range of populations and exposures. Over the years, the limits have been lowered considerably: for example, the annual limit of the occupational effective dose was 150 mSv in 1950 and stands at 20 mSv in 2011 (100 mSv averaged over 5 years, with doses not exceeding 50 mSv in any single year).

Internationally, the United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) was established by the General Assembly of the UN in 1955 amidst growing concerns about effects of radiation. UNSCEAR assesses and reports levels and effects of exposure to ionizing radiation worldwide. Regularly, all relevant epidemiological evidence is summarized and evaluated. The committee’s publications form the basis for evaluation of radiation risk and the development of radiation protection policies, notably by the ICRP (UNSCEAR 2010).

In the 1960s, it became clear that the genetic effect had probably been overestimated in some of the older studies but that cancer mortality and incidence risk increases were now apparent, namely, among atomic bomb survivors. These observations resulted in substantial changes in the discussion on radiation risks and protection. The focus was now on exposure to low doses of ionizing radiation as there was no suggestion of the existence of a threshold for late health effects, notably for cancer.

In practical terms, the ICRP also proposed a philosophy of protection by defining the three principles of justification, optimization of protection, and dose limitation which was confirmed recently (ICRP 2007) (see Box 50.2). These principles as well as other guidance from the ICRP have been adopted by national radiation protection authorities worldwide.

Box 50.2. Principles of radiation protection

Justification: the benefits of using radiation must outweigh the negative aspects.
Optimization: radiation exposure must be kept as low as reasonable achievable (ALARA principle).
Limitation: doses to individuals must not exceed the recommended dose limits.

Radiation protection covers exposures resulting not only from the operation of radiation sources ("planned exposure situations") but also from emergency exposure situations, for instance resulting from nuclear accidents, and a range of other situations such as those involving exposure to natural radiation sources ("existing exposure situations") (Wrixon 2008). Data from epidemiology continue to be an important scientific basis for all radiation protection efforts.

50.7 Conclusions

Radiation epidemiology has developed over many years to become an area where both exposure assessment and analytical epidemiological methods have reached a considerable level of maturity. Radiation protection standards thus have a firm foundation in epidemiology, with studies on a large variety of populations and exposure situations. Nevertheless, open questions remain. These questions refer particularly to the dose levels relevant for the general public, namely, the doses below about 100 mSv, where the statistical power of studies becomes low and the potential for biased results increases. Such low doses are often delivered at low dose rates as well, and there is some controversy whether high dose rates (e.g., from acute accidental exposures) or low dose rates lead to similar risk increases. Multidisciplinary approaches combining experimental radiobiology and radiation epidemiology as well as other disciplines are seen as an option to better understand and quantify risks at low doses.

In terms of exposure assessment, the retrospective quantification of ionizing radiation exposures continues to be challenging. New or extended cohorts and other studies will probably benefit from improving documentation and the growing availability of electronic records of exposures over time, for example, in occupational and medical settings. The investigation of cancer risks from diagnostic CT exposures is an example of such an approach and also highlights the fact that medical exposures are of major relevance for the overall population dose and thus for radiation epidemiology and radiation protection. Biological markers of radiation exposure will play a role in studies where biological material is available, but reliable and stable biomarkers especially for the low dose range are needed.

Another group of radiation effects, the so-called non-targeted effects of radiation, as opposed to the effects directly induced by the energy deposition of radiation in cells, has been a major research issue in experimental radiation biology during recent years, with little evidence from human epidemiological studies so far (UNSCEAR 2006, Annex C). Genomic instability refers to an increased rate of genomic alterations induced by ionizing radiation, which may lead to effects in later cell generations. A related concept is the so-called bystander effect describing effects in cells that are not directly hit by radiation. The relevance of these non-targeted effects for radiation epidemiology and radiation risk assessment remains inconclusive at present. Interdisciplinary research efforts focusing, inter alia, on genetic and epigenetic aspects of radiation risk are under way and will possibly challenge established concepts in the field (Averbeck 2010).

In terms of outcomes, radiation epidemiology has focused on cancer as a late effect of exposure to ionizing radiation in the past, while other health consequences received less attention. This is now changing, with increased research efforts on cardiovascular diseases, lens opacities, and cognitive development as outcomes of interest in radiation epidemiology. Some of these issues can be studied in established cohorts, whereas specific groups such as interventional radiologists or cardiologists may be included in new cohorts, for example, to assess cataract risks.

Recent events such as the Fukushima nuclear accident underline the need for ongoing research efforts as the basis for risk assessment and communication with the public and others. Despite the comparatively large body of evidence, health effects of ionizing radiation are often misrepresented or misunderstood. Radiation epidemiology supports radiation risk assessment with increasingly detailed and methodologically advanced scientific information. As radiation risks are dreaded by the public, a particular challenge remains in the approaches to translate concepts and results of radiation epidemiology into information suitable for risk communication in public health.

References

Ainsbury EA, Bouffler SD, Dörr W, Graw J, Muirhead CR, Edwards AA, Cooper J (2009) Radiation cataractogenesis: a review of recent studies. Radiat Res 172:1–9

Ashmore JP, Gentner NE, Osborne RV (2010) Incomplete data on the Canadian cohort may have affected the results of the study by the IARC on the radiogenic cancer risk among nuclear industry workers in 15 countries. J Radiol Prot 30:121–129

Averbeck D (2010) Non-targeted effects as a paradigm breaking evidence. Mutat Res 687:7–12

Bartlett DT (2004) Radiation protection aspects of the cosmic radiation exposure of aircraft crew. Radiat Prot Dosim 109:349–355

Berrington A, Darby SC, Weiss HA, Doll R (2001) 100 years of observation on British radiologists: mortality from cancer and other causes 1897–1997. Br J Radiol 74:507–519

Blettner M, Zeeb H, Auvinen A, Ballard TJ, Caldora M, Eliasch H, Gundestrup M, Haldorsen T, Hammar N, Hammer GP, Irvine D, Langner I, Paridou A, Pukkala E, Rafnsson V, Storm H, Tulinius H, Tveten U, Tzonou A (2003) Mortality from cancer and other causes among male airline cockpit crew in Europe. Int J Cancer 106:946–952

Brenner DJ, Elliston CD (2004) Estimated radiation risks potentially associated with full-body CT screening. Radiology 232:735–738

Brenner DJ, Elliston CD, Hall EJ, Berdon WE (2001) Estimated risks of radiation-induced fatal cancer from pediatric CT. Am J Roentgenol 176:289–296

Boice JD (1981) Cancer following medical irradiation. Cancer 47:1081–1090

Boice JD, Lubin JH (1997) Occupational and environmental radiation and cancer. Cancer Causes Control 8:309–322

Cardis E, Hatch M (2011) The Chernobyl accident – an epidemiological perspective. Clin Oncol (R Coll Radiol) 23:251–260

Cardis E, Vrijheid M, Blettner M, Gilbert E, Hakama M, Hill C, Howe G, Kaldor J, Muirhead CR, Schubauer-Berigan M, Yoshimura T, Bermann F, Cowper G, Fix J, Hacker C, Heinmiller B, Marshall M, Thierry-Chef I, Utterback D, Ahn Y-O, Amoros E, Ashmore P, Auvinen A, Bae J-M, Solano Bernar J, Biau A, Combalot E, Deboodt P, Diez Sacristan A, Eklof M, Engels H, Engholm G, Gulis G, Habib R, Holan K, Hyvonen H, Kerekes A, Kurtinaitis J, Malker H, Martuzzi M, Mastauskas A, Monnet A, Moser M, Pearce MS, Richardson DB, Rodriguez-Artalejo F, Rogel A, Tardy H, Telle-Lamberton M, Turai I, Usel M, Veress K (2005) Risk of cancer after low doses of ionising radiation: retrospective cohort study in 15 countries. BMJ 331:77

Cardis E, Vrijheid M, Blettner M, Gilbert E, Hakama M, Hill C, Howe G, Kaldor J, Muirhead CR, Schubauer-Berigan M, Yoshimura T, Bermann F, Cowper G, Fix J, Hacker C, Heinmiller B, Marshall M, Thierry-Chef I, Utterback D, Ahn Y-O, Amoros E, Ashmore P, Auvinen A, Bae J-M, Bernar J, Biau A, Combalot E, Deboodt P, Diez Sacristan A, Eklöf M, Engels H, Engholm G, Gulis G, Habib R, Holan K, Hyvonen H, Kerekes A, Kurtinaitis J, Malker H, Martuzzi M, Mastauskas A, Monnet A, Moser M, Pearce MS, Richardson DB, Rodriguez-Artalejo F, Rogel A, Tardy H, Telle-Lamberton M, Turai I, Usel M, Veress K (2007) The 15-country collaborative study of cancer risk among radiation workers in the nuclear industry: estimates of radiation-related cancer risks. Radiat Res 167:396–416

Cullings HM, Fujita S, Funamoto S, Grant EJ, Kerr GD, Preston DL (2006) Dose estimation for atomic bomb survivor studies: its evolution and present status. Radiat Res 166:219–254

Darby S, Hill D, Auvinen A, Barros-Dios JM, Baysson H, Bochicchio F, Deo H, Falk R, Forastiere F, Hakama M, Heid I, Kreienbrock L, Kreuzer M, Lagarde F, Mäkeläinen I, Muirhead C, Oberaigner W, Pershagen G, Ruano-Ravina A, Ruosteenoja E, Schaffrath Rosario A, Tirmarche M, Tomásek L, Whitley E, Wichmann HE, Doll R (2005) Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. BMJ 330:223

de Vathaire F, Francois P, Hill C, Schweisguth O, Rodary C, Sarrazin D, Oberlin O, Beurtheret C, Dutreix A, Flamant R (1989) Role of radiotherapy and chemotherapy in the risk of second malignant neoplasms after cancer in childhood. Br J Cancer 59:792–796

Doll R, Wakeford R (1997) Risk of childhood cancer from fetal irradiation. Br J Radiol 70:130–139

Drozdovitch V, Bouville A, Chobanova N, Filistovic V, Ilus T, Kovacic M, Malátová I, Moser M, Nedveckaite T, Völkle H, Cardis E (2007) Radiation exposure to the population of Europe following the Chernobyl accident. Radiat Prot Dosim 123:515–528

Eidemüller M, Ostroumova E, Krestinina L, Epiphanova S, Akleyev A, Jacob P (2010) Comparison of mortality and incidence solid cancer risk after radiation exposure in the Techa River Cohort. Radiat Environ Biophys 43:477–490

Federal Office for Radiation Protection (2012) Frequently Asked Questions (FAQs) Ionising Radiation. http://www.bfs.de/en/ion/faq/faq_istrahlung.html/#4. Accessed 26 Feb 2012

Frieben A (1902) Demonstration eines Cancroids des rechten Handrückens, das sich nach langdauernder Einwirkung von Roentgenstrahlen entwickelt hatte. Fortschr Roentgenstr 6:106–111

Gilman EA, Steward AM, Knox EG, Kneale GW (1989) Trends in obstetric radiography, 1939–1981. J Radiol Prot 9:93–101

Greaves M (2006) Infection, immune responses and the etiology of childhood leukemia. Nat Rev Cancer 6:193–203

Grosche B, Kreuzer M, Kreisheimer M, Schnelzer M, Tschense A (2006) Lung cancer risk among German male uranium miners: a cohort study, 1946–1998. Br J Cancer 95:1280–1287

Grupen C (2010) Introduction to radiation protection: practical knowledge for handling radioactive sources. Springer, Berlin/Heidelberg

Hall EJ, Brenner DJ (2008) Cancer risks from diagnostic radiology. Br J Radiol 81:362–378

Hammer GP, Blettner M, Zeeb H (2009a) Epidemiological studies of cancer in aircrew. Radiat Prot Dosim 136:232–239

Hammer GP, Seidenbusch M, Schneider K, Regulla DF, Zeeb H, Spix C, Blettner M (2009b) A cohort study of childhood cancer incidence after postnatal diagnostic X-ray exposure. Radiat Res 171:504–512

Häusler U, Czarwinski R, Brix G (2009) Radiation exposure of medical staff from interventional x-ray procedures: a multicentre study. Eur Radiol 19:2000–2008

Hawkins MM, Draper GJ, Kingston JE (1987) Incidence of second primary tumours among childhood cancer survivors. Br J Cancer 56:339–347

Hendry JH, Simon SL, Wojcik A, Sohrabi M, Burkart W, Cardis E, Laurier D, Tirmarche M, Hayata I (2009) Human exposure to high natural background radiation: what can it teach us about radiation risks? J Radiol Prot 29:A29–A42

Hesse O (1911) Symptomatologie, Pathogenese und Therapie des Röntgenkarzinoms. JA Barth, Leipzig

Howe GR, McLaughlin J (1996) Breast cancer mortality between 1950 and 1987 after exposure to fractionated moderate-dose-rate ionizing radiation in the Canadian fluoroscopy cohort study and a comparison with breast cancer mortality in the atomic bomb survivors study. Radiat Res 145:694–707

IARC (2000) Monographs on the evaluation of carcinogenic risks to humans. Ionizing radiation, part 1: X- and gamma (y)-radiation, and neutrons. IARC Monograph, vol 75. IARC, Lyon

ICRP (2007) The 2007 Recommendations of the International Commission on Radiological Protection. ICRP Publ. 103, Ann ICRP 37(2–4):1–332

ICRP (2009) Application of the commission's recommendations for the protection of people in emergency exposure situations. ICRP Publ. 109, Ann ICRP 39(1):1–110

Ivanov VK, Chekin SY, Kashcheev VV, Maksioutov MA, Tumanov KA (2008) Risk of thyroid cancer among Chernobyl emergency workers of Russia. Radiat Environ Biophys 47:463–467

Jacob V, Jacob P (2004) Modelling of carcinogenesis and low-dose hypersensitivity: an application to lung cancer incidence among atomic bomb survivors. Radiat Environ Biophys 42:265–273

Jacob P, Ruhm W, Walsh L, Blettner M, Hammer G, Zeeb H (2009) Is cancer risk of radiation workers larger than expected? Occup Environ Med 66:789–796

Kaatsch P, Spix C, Schulze-Rath R, Schmiedel S, Blettner M (2008) Leukaemia in young children living in the vicinity of German nuclear power plants. Int J Cancer 1220:721–726

Kim KP, Lee J, Bolch WE (2011) CT dosimetry computer codes: their influence on radiation dose estimates and the necessity for their revision under new ICRP radiation protection standards. Radiat Prot Dosim 146:252–255

Kleinerman RA (2006) Cancer risks following diagnostic and therapeutic radiation exposure in children. Pediatr Radiol 36:121–125

Krestinina LY, Davis F, Ostroumova E, Epifanova S, Degteva M, Preston D, Akleyev A (2007) Solid cancer incidence and low-dose-rate radiation exposures in the Techa River cohort: 1956–2002. Int J Epidemiol 36:1038–1046

Lightfoot TJ, Roman E (2004) Causes of childhood leukaemia and lymphoma. Toxicol Appl Pharmacol 199:104–117

Linet MS, Freedman DM, Mohan AK (2005) Incidence of haematopoetic malignancies in US radiologic technologists. Occup Environ Med 62:861–867

Lubin JH, Boice JD, Edling C, Hornung RW, Howe GR, Kunz E, Kusiak RA, Morrison HI, Radford EP, Samet JM, Tirmarche M, Woodward A, Yao SX, Pierce DA (1995) Lung cancer in radon-exposed miners and estimation of risk from indoor exposure. J Natl Cancer Inst 87:817–827

March HC (1944) Leukaemia in radiologists. Radiology 43:275–278

Martland HS, Humphries RE (1929) Osteogenic sarcoma in dial painters using luminous paint. Arch Pathol 7:406–417

Michel R, Handl J, Ernst T, Botsch W, Szidat S, Schmidt A, Jakob D, Beltz D, Romantschuk LD, Synal HA, Schnabel C, López-Gutiérrez JM (2005) Iodine-129 in soils from Northern Ukraine and the retrospective dosimetry of the iodine-131 exposure after the Chernobyl accident. Sci Total Environ 340:35–55

Muirhead CR, O'Hagan JA, Haylock RGE, Phillipson MA, Willcock T, Berridge GL, Zhang W (2009) Mortality and cancer incidence following occupational radiation exposure: third analysis of the national registry for radiation workers. Br J Cancer 100:206–212

National Research Council (2006) Health risks from exposures to low levels of ionizing radiation BEIR VII Phase 2. The National Academies Press, Washington DC

Ozasa K, Shimizu Y, Suyama A, Kasagi F, Soda M, Grant EJ, Sakata R, Sugiyama H, Kodama K (2012) Studies of the mortality of atomic bomb survivors, Report 14, 1950–2003: An overview of cancer and noncancer diseases. Radiat Res 177:229–243

Parkin DM, Darby SC (2011) Cancers in 2010 attributable to ionising radiation exposure in the UK. Br J Cancer 105:57–65

Parkin DM, Stiller CA, Draper GJ, Bieber CA (1988) The international incidence of childhood cancer. Int J Cancer 42(4):511–520

Pearce MS, Salotti JA, Little MP, McHugh K, Lee C, Kim KP, Howe NL, Ronckers CM, Rajaraman P, Craft AW, Parker L, Berrington de González A (2012) Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. Lancet 380:499–505. Jun 7 (Epub 2012 June 7)

Pinto MM, Santos NF, Amaral A (2010) Current status of biodosimetry based on standard cytogenetic methods. Radiat Environ Biophys 49:567–581

Preston DL, Shimizu Y, Pierce DA, Suyama A, Mabuchi K (2003) Studies of mortality of atomic bomb survivors. Report 13: solid cancer and noncancer disease mortality: 1950–1997. Radiat Res 160:381–407

Preston DL, Pierce DA, Shimizu Y, Cullings HM, Fujita S, Funamoto S, Kodama K (2004) Effect of recent changes in atomic bomb survivor dosimetry on cancer mortality risk estimates. Radiat Res 162:377–389

Preston DL, Ron E, Tokuoka S, Funamoto S, Nishi N, Soda M, Mabuchi K, Kodama K (2007) Solid cancer incidence in atomic bomb survivors: 1958–1998. Radiat Res 168:1–64

Reulen RC, Frobisher C, Winter DL, Kelly J, Lancashire ER, Stiller CA, Pritchard-Jones K, Jenkinson HC, Hawkins MM, British Childhood Cancer Survivor Study Steering Group (2011) Long-term risks of subsequent primary neoplasms among survivors of childhood cancer. JAMA 305:2311–2319

Romanenko A, Bebesho V, Hatch M, Bazyka D, Finch S, Dyagil I, Reiss R, Chumak V, Bouville A, Gudzenko N, Zablotska L, Pilinskaya M, Lyubarets T, Bakhanova E, Babkina N, Trotsiuk N, Ledoschuk B, Belayev Y, Dybsky SS, Ron E, Howe G (2008) The Ukrainian-American study of Leukemia and related disorders among Chernobyl cleanup workers from Ukraine: I. Study methods. Radiat Res 170:691–697

Rühm W (2003) Neutronendosimetrie in Hiroshima. Phys J 2:37–42

Sermage-Faure C, Laurier D, Goujon-Bellec S, Chartier M, Guyot-Goubin A, Rudant J, Hémon D, Clavel J (2012) Childhood leukemia around French nuclear power plants – the Geocap study, 2002–2007. Int J Cancer 131:E769–E780

Shilnikova NS, Preston DL, Ron E, Gilbert ES, Vassilenko EK, Romanov SA, Kuznetsova IS, Sokolikov ME, Okatenko PV, Kreslov VV, Koshurnikova NA (2003) Cancer mortality risk among workers at the Mayak nuclear complex. Radiat Res 159:787–798

Shimizu Y, Kodama K, Nishi N, Kasagi F, Suyama A, Soda M, Grant EJ, Sugiyama H, Sakata R, Moriwaki H, Hayashi M, Konda M, Shore RE (2010) Radiation exposure and circulatory disease risk: Hiroshima and Nagasaki atomic bomb survivor data, 1950–2003. BMJ 340:b5349

Shore RE, Hildreth N, Woodard E, Dvoretsky P, Hempelmann L, Pasternack B (1986) Breast cancer among women given x-ray therapy for acute postpartum mastitis. J Natl Cancer Inst 77:689–696

Sigurdson AJ, Ron E (2004) Cosmic radiation exposure and cancer risk among flight crew. Cancer Invest 22:743–761

Sigurdson AJ, Doody MM, Rao RS, Freedman DM, Alexander BH, Hauptmann M, Mohan AK, Yoshinage S, Hill DA, Tarone R, Mabuchi K, Ron E, Linet MS (2003) Cancer incidence in the US radiologic technologists health study, 1983–1998. Cancer 97:3080–3089

Simon SL, Weinstock RM, Doody MM, Neton J, Wenzl T, Stewart P, Mohan AK, Yoder RC, Hauptmann M, Freedman DM, Cardarelli J, Feng HA, Bouville A, Linet M (2006) Estimating historical radiation doses to a cohort of U.S. radiologic technologists. Radiat Res 166:174–192

Sokolnikov ME, Gilbert ES, Preston DL, Ron E, Shilnikova NS, Khokhryakov VV, Vasilenko EK, Koshurnikova NA (2008) Lung, liver and bone cancer mortality in Mayak workers. Int J Cancer 123:905–911

Spycher BD, Feller M, Zwahlen M, Röösli M, von der Weid NX, Hengartner H, Egger M, Kuehni CE, Swiss Paediatric Oncology Group, Swiss National Cohort Study Group (2011) Childhood cancer and nuclear power plants in Switzerland: a census-based cohort study. Int J Epidemiol 40:1247–1260

Stamm G, Nagel HD (2002) CT-Expo – ein neuartiges Programm zur Dosisevaluierung in der CT. Fortschr Röntgenstr 174:1570–1576

Stebbings JH, Lucas HF, Stehney AF (1984) Mortality from cancers of major sites in female radium dial workers. Am J Ind Med 5:435–459

Steward A, Webb J, Giles D, Hewitt D (1956) Malignant disease in childhood and diagnostic irradiation in utero. Lancet 2:447

Steward A, Webb J, Hewitt D (1958) A survey of childhood malignancies. Br Med J 1:1495–1508

Thierry-Chef I, Marshall M, Fix JJ, Bermann F, Gilbert ES, Hacker C, Heinmiller B, Murray W, Pearce MS, Utterback D, Bernar K, Deboodt P, Eklof M, Griciene B, Holan K, Hyvonen H, Kerekes A, Lee M-C, Moser M, Pernicka F, Cardis E (2007) The 15-country collaborative study of cancer risk among radiation workers in the nuclear industry: study of errors in dosimetry. Radiat Res 167:380–395

Tucker MA, Morris Jones PH, Boice JD Jr, Robison LL, Stone BJ, Stovall M, Jenkin RD, Lubin JH, Baum ES, Siegel SE, Meadows AT, Hoover RN, Fraumeni JF Jr (1991) Therapeutic radiation at a young age is linked to secondary thyroid cancer. Cancer Res 51:2885–2888

UNSCEAR, United Nations Scientific Committee on the Effects of Atomic Radiation (2000) Sources and effects of ionizing radiation. Vol II. United Nations, New York

UNSCEAR, United Nations Scientific Committee on the Effects of Atomic Radiation (2006) Effects of ionizing radiation. Vol II, Annex C. United Nations, New York

UNSCEAR, United Nations Scientific Committee on the Effects of Atomic Radiation (2008) Effects of ionizing radiation. Vol II, Annex D. United Nations, New York

UNSCEAR, United Nations Scientific Committee on the Effects of Atomic Radiation (2010) Summary of low-dose radiation effects on health. United Nations, New York

Upton AC (1977) Radiobiological effects of low doses: implications for radiological protection. Radiat Res 71:51–54

Vasilenko EK, Khokhryakov VF, Miller SC, Fix JJ, Eckerman K, Choe DO, Gorelov M, Khokhryakov VV, Knyasev V, Krahenbuhl MP, Scherpelz RI, Smetanin M, Suslova K, Vostrotin V (2007) Mayak worker dosimetry study: an overview. Health Phys 93:190–206

Vrijheid M, Cardis E, Blettner M, Gilbert E, Hakama M, Hill C, Howe G, Kaldor J, Muirhead CR, Schubauer-Berigan M, Yoshimura T, Ahn Y-O, Ashmore P, Auvinen A, Bae J-M, Engels H, Gulis G, Habib RR, Hosoda Y, Kurtinaitis J, Malker H, Moser M, Rodriguez-Artalejo F, Rogel A, Tardy H, Telle-Lamberton M, Turai I, Usel M, Veress K (2007) The 15-country collaborative study of cancer risk among radiation workers in the nuclear industry: design, epidemiological methods and descriptive results. Radiat Res 167:361–379

Wakeford R (2005) Cancer risk among nuclear workers. J Radiol Prot 25(3):225–228

Wakeford R (2009) Radiation in the workplace – a review of studies of the risk of occupational exposure to ionising radiation. J Radiol Prot 29:A61–A79

Wakeford R, Little MP (2003) Risk coefficients for childhood cancer after intrauterine irradiation: a review. Int J Radiat Biol 79:293–309

Walsh L, Tschense A, Schnelzer M, Dufey F, Grosche B, Kreuzer M (2010) The influence of radon exposures on lung cancer mortality in German uranium miners, 1946–2003. Radiat Res 173:79–90

WHO, World Health Organization (2009) WHO handbook on indoor radon: a public health perspective. World Health Organization, Geneva

Wrixon AD (2008) New ICRP recommendations. J Radiol Prot 28:161–168

Zeeb H, Blettner M, Langner I, Hammer GP, Ballard TJ, Santaquilani M, Gundestrup M, Storm H, Haldorsen T, Tveten U, Hammar N, Linnersjö A, Velonakis E, Tzonou A, Auvinen A, Pukkala E, Rafnsson V, Hrafnkelsson J (2003) Mortality from cancer and other causes among airline cabin attendants in Europe: a collaborative cohort study in eight countries. Am J Epidemiol 158:35–46

Part V

Outcome-Oriented Epidemiology

Infectious Disease Epidemiology

51

Susanne Straif-Bourgeois, Raoult Ratard,
and Mirjam Kretzschmar

Contents

S. Straif-Bourgeois (✉) • R. Ratard
Louisiana Department of Health and Hospitals, Office of Public Health, Infectious Disease Epidemiology Section, New Orleans, LA, USA

M. Kretzschmar
Julius Centre for Health Sciences and Primary Care University Medical Centre Utrecht, CX Utrecht, The Netherlands

Centre for Infectious Disease Control, RIVM, MA Bilthoven, The Netherlands

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_34,

51.1 Introduction

51.1.1 Specifics of Infectious Disease Epidemiology

Most textbooks dealing with the epidemiology of infectious diseases address the epidemiological features (also named biology) of specific infectious diseases. In this chapter, the focus is placed on the concepts and methods more specific to the general epidemiological study of infectious diseases. At a later stage, implementation of these methods must be adapted to the specific infectious disease under consideration. Then, detailed knowledge of the disease biology is of capital importance.

Epidemiologists focus their study on population groups (or "herds") rather than on individuals. In addition, infectious disease epidemiology also considers the interaction between individuals within the population group. For non-infectious diseases, each case and his/her risk factors are personal and independent from the neighbor (your neighbor's risk factors for heart disease have no influence on your risk factors). On the contrary, for infectious disease, the interaction between cases and contacts is of prime importance; this special feature of infectious disease epidemiology is discussed in the section transmission and basic concepts important to infectious diseases.

Although not entirely specific of infectious disease epidemiology, some characteristics are more often found in this field of study; for example, infectious disease epidemiology is:

- The closest to "shoe leather" epidemiology, meaning going into the community, talking to patients, contacts, practitioners, observing the environment (living conditions, activities, food preparation, water supply, etc.)
- Direct understanding and "closeness" to data
- Small-scale investigations
- Immediate results
- Easy understanding of etiology

51.1.2 The Global Burden of Infectious Diseases

Infectious diseases are a major cause of human suffering in terms of both morbidity and mortality throughout human history. The spread of infectious diseases was influenced by various steps in human civilization. For example, parasitic and zoonotic diseases have become more common after domestication of animals, airborne viral and bacterial infections after large settlements and urbanization. Throughout the ages, humanity suffered from large pandemics such as plague, smallpox, cholera, and influenza but also from the more silent killers of chronic infectious diseases such as tuberculosis and syphilis.

Morbidity due to infectious diseases is very common in spite of the progress accomplished in recent decades. According to the World Health Organization's (WHO) annual estimates, there are globally 300–500 million cases of malaria, 333 million cases of sexually transmitted diseases (syphilis, gonorrhea, chlamydia,

Table 51.1 Top ten causes of death worldwide (WHO 2008)

	Deaths in millions per year	% of all deaths
Ischemic heart disease	7.25	12.8
Stroke and other cerebrovascular disease	6.15	10.8
Lower respiratory infections	3.46	6.1
Chronic obstructive pulmonary disease	3.28	5.8
Diarrheal diseases	2.46	4.3
HIV/AIDS	1.78	3.1
Trachea, bronchus, lung cancers	1.39	2.4
Tuberculosis	1.34	2.4
Diabetes mellitus	1.26	2.2
Road traffic accidents	1.21	2.1

and trichomonas), 33 million cases of HIV/AIDS, 14 million people infected with tuberculosis, and 3–5 million cases of cholera (WHO 2010).

Even though infectious diseases are much more common in the non-industrialized world, the prevalence of infection is still very high for some infectious diseases in the industrialized world. Annually, approximately 48 million episodes of diarrhea are leading to 128,000 hospitalizations, and 3,000 deaths due to diarrheal illnesses are occurring in the United States (Centers for Disease Control and Prevention (CDC) 2011; Mounts et al. 1999). Every year, influenza virus circulates widely, infecting from 10% to 40% of the world population. Based on CDC estimates, there were 59 million infected during the 2009/2010 H1N1 pandemic (CDC 2011). Furthermore, serological surveys found that by young adulthood, the prevalence of antibodies was 80% against herpes simplex virus type 1, 15–20% against type 2, 95% against human herpes virus, 63% against *Hepatitis A*, 2% against *Hepatitis C*, 0.5% against *Hepatitis B*, and 50% against *Chlamydia pneumoniae* (American Academy of Pediatrics 2006; Mandell et al. 2000).

Not surprisingly, there is also a large imbalance in mortality rates due to infectious diseases between non-industrialized and industrialized countries. Globally, every third death is due from an infectious disease. In 1990, estimated 17 million deaths were due to communicable diseases, along with malnutrition and maternal and perinatal diseases with about 95% of these deaths occurring in the poorest parts of the world, mainly India and sub-Saharan continent (University of California 2011). According to the WHO, the most common causes of infectious disease deaths were lower respiratory infections (3.46 million), diarrheal diseases (2.46 million), HIV/AIDS (1.78 million), tuberculosis (1.34), malaria (1.1 million), and measles (900,000) (WHO 2008) (see Table 51.1).

51.1.3 The Importance of Infectious Disease Epidemiology for Prevention

It is often said that "Epidemiology is the basic science of preventive medicine." To prevent diseases, it is important to understand the causative agents, risk factors, and circumstances that lead to a specific disease. This is even more important for infectious disease prevention, since simple interventions may break the chain of transmission. Preventing cardiovascular diseases or cancer is much more difficult because it usually requires multiple long-term interventions requiring lifestyle changes and behavior modification, which are difficult to achieve.

In 1900, the American Commission of Yellow Fever, headed by Walter Reed, was sent to Cuba. The commission showed that the infective agent was transmitted by the mosquito *Aedes aegypti*. This information was used by the then Surgeon General of the US Army, William Gorgas, to clean up the 200-year-old focus of yellow fever in Havana by using mosquito proofing or oiling of the larval habitat, dusting houses with pyrethrum powder, and isolating suspects under a mosquito net. This rapidly reduced the number of cases in Havana from 310 in 1900 to 18 in 1902 (Goodwin and Gordon Smith 1996).

A complete understanding of the causative agent and transmission is always useful but not absolutely necessary. The most famous example is that of John Snow who was able to link cholera transmission to water contamination during the London cholera epidemic of 1854 by comparing the deaths from those households served by the Southwark and Vauxhall Company versus those served by another water company. John Snow further confirmed his hypothesis by the experiment of removing the Broad Street pump handle (Wills 1996a).

51.1.4 The Changing Picture of Infectious Disease Epidemiology

Over the past three decades, more than 40 new pathogens have been identified, some of them with global importance: *Bartonella henselae*, *Borrelia burgdorferi*, *Campylobacter*, *Cryptosporidium*, *Cyclospora*, *ebola virus*, *Escherichia coli* 0157:H7, *Ehrlichia*, *hantaan virus*, *Helicobacter*, *hendra virus*, *Hepatitis C* and *E*, *HIV*, *HTLV-I* and *II*, *human herpes virus 6* and *8*, *human metapneumovirus*, *Legionella*, new variant Creutzfeldt-Jakob disease agent, *nipah virus*, *norovirus*, *Parvovirus B19*, *rotavirus*, severe acute respiratory syndrome (SARS), etc.

While there are specific causative agents for infectious diseases, these agents may undergo some changes over time. The last major outbreak of pneumonic plague (*Yersinia pestis*) in the world occurred in Manchuria in 1921. This scourge, which had decimated humans for centuries, is no longer a major threat. The plague bacillus cannot survive long outside its animal host (humans, rodents, fleas) because it lost the ability to complete the Krebs cycle on its own. While it can only survive in its hosts, the plague bacillus also destroys its hosts rapidly. As long as susceptible hosts were abundant, plague did prosper. When environmental conditions became less favorable (lesser opportunities to sustain the host to host cycles), less virulent strains had a selective advantage (Wills 1996b).

51.1.4.1 Changes in Etiological Agent

The influenza virus is the best example of an agent able to undergo changes leading to renewed ability to infect populations that had been already infected and immune. The influenza virus is a single-stranded RNA virus with a lipophilic envelope. Two important glycoproteins from the envelope are the hemagglutinin (HA) and neuraminidase (NA). The HA protein is able to agglutinate red blood cells (hence its name). This protein is important as it is a major antigen for eliciting neutralizing antibodies. *Antigenic drift* is a minor change in surface antigens that result from point mutations in a gene segment. Antigenic drift may result in epidemics, since incomplete protection remains from past exposures to similar viruses. *Antigenic shift* is a major change in one or both surface antigens (H and/or N) that occurs at varying intervals. Antigenic shifts are probably due to genetic recombination (an exchange of a gene segment) between influenza type A viruses, usually those that affect humans and birds. An antigenic shift may result in a worldwide pandemic if the virus can be efficiently transmitted from person to person.

51.1.4.2 Changes in Populations at Risk

In the past three decades throughout the world, there has been a shift toward an increase in the population of individuals at high risk for infectious diseases. In industrialized nations, the increase in longevity leads to higher proportion of the elderly population who are more prone to acquiring infectious diseases and developing life-threatening complications. For example, a West Nile virus (WNV) infection is usually asymptomatic or causes a mild illness (West Nile fever); rarely does it cause a severe neuroinvasive disease. In the 2002 epidemic of West Nile virus in Louisiana, the incidence of neuroinvasive disease increased progressively from 0.3 per 100,000 in the 0 to 14 age group to 9 per 100,000 in the 60- to 75-year-old age group and jumped to 32 per 100,000 in the age group 75 and older. Mortality rates showed the same pattern, a gradual increase to 0.7 per 100,000 in the 60 to 75 age group with a sudden jump to 11 per 100,000 for the oldest age group of 75 and older (Balsamo et al. 2003).

Improvement in health care in industrialized nations has caused an increase in the number of immune-deficient individuals, be it cancer survivors, transplant patients, or people on immunosuppressive drugs for long-term autoimmune diseases. Some of the conditions that may increase susceptibility to infectious diseases are cancers, particularly patients on chemo- or radiotherapy, leukemia, lymphoma, Hodgkin's disease, immune suppression (HIV infection), long-term steroid use, liver disease, hemochromatosis, diabetes, alcoholism, chronic kidney disease, and dialysis patients. For example, persons with liver disease are 80 times more likely to develop *Vibrio vulnificus* infections than are persons without liver disease. Some of these infections may be severe, leading to death.

In developing countries, a major shift in population susceptibility is associated with the high prevalence of immune deficiencies due to HIV infections and AIDS. In Botswana, which has a high prevalence of HIV (sentinel surveillance revealed HIV seroprevalence rates of 36% among women presenting for routine antenatal care), tuberculosis rates increased from 202 per 100,000 in 1989 to 537 per 100,000 in 1999 (Lockman et al. 2001), while before the HIV/AIDS epidemics, rates above 100 were very rare.

Changes in lifestyles have increased opportunities for the transmission of infectious disease agents in populations previously at low risk. Intravascular drug injections have increased the transmission of agents present in blood and body fluids (e.g., HIV, Hepatitis B and C). Consumption of raw fish, shellfish, and ethnic food expanded the area of distribution of some parasitic diseases. Air travel allows people to be infected in a country and be halfway around the globe before becoming contagious.

By the same token, insects and other vectors have become opportunistic global travelers. *Aedes albopictus*, the Asian tiger mosquito, which is the vector for dengue, eastern equine encephalitis, and other viruses, was thus imported in 1985 to Houston, Texas, inside Japanese tires. Subsequently, it has invaded 26 US states.

51.1.4.3 Changes in Knowledge About Transmission of Disease Agents

With the advent of nucleic acid tests, it has become possible to detect the presence of infectious disease agents in the air and environmental surfaces. For example, the use of air samplers and polymerase chain reaction analysis has shown that *Bordetella pertussis* DNA can be found in the air surrounding patients with *B. pertussis* infection, providing further evidence of airborne spread (Aintablian et al. 1998) and thus leading to reevaluate the precautions to be taken. However, the presence of nucleic acids in an environmental medium does not automatically mean that transmission will occur. Further studies are necessary to determine the significance of such findings.

51.1.4.4 Bioterrorism Adds a New Dimension

Infectious disease agents, when used in bioterrorism events, have often been reengineered to have different physical properties and are used in quantities not usually experienced in natural events. There is little experience and knowledge about the human body's response to large doses of an infectious agent inhaled in aerosol particles that are able to be inhaled deep into lung alveolae. The probably best examples are the 2001 anthrax attacks in the United States. One week after the September 11 attacks, letters containing anthrax spores were mailed to several news media offices and two US Senators, killing a total of 5 persons and infecting 17 others. During the course of these anthrax letter events, there was considerable discussion about incubation period, recommended duration of prophylaxis, and minimum infectious dose for aerosolized and reengineered anthrax spores. The lack of knowledge base has led to confusion in recommendations being made.

51.2 New Approaches

Although the basics of infectious disease epidemiology have not changed and the discipline remains strongly anchored on some basic principles, technological developments such as improved laboratory methods and enhanced use of informatics (such as advanced mapping tools, web-based reporting systems, and statistical analytical software) have greatly expanded the field of infectious disease epidemiology.

51.2.1 Improved Laboratory Methods

Molecular techniques are being used more and more as a means to analyze epidemiological relationships between microorganisms. Hence, the term molecular epidemiology refers to epidemiological research studies made at the molecular level (see also chapter ▸Molecular Epidemiology of this handbook).

The main microbial techniques use target plasmids and chromosomes, more specifically, plasmid fingerprinting and plasmid restriction endonuclease (REA) digestion, chromosomal analysis including pulse field gel electrophoresis (PFGE), restriction fragment length polymorphism (RFLP), multi-locus sequence type (MLST), and spa typing to name a few of these techniques. Polymerase chain reaction (PCR) is used to amplify the quantity of genomic material present in the specimen. Real-time PCR detection of infectious agents is now possible in a few hours. These techniques are becoming more widely used, even in public health laboratories for routine investigations. For more detailed information on these molecular techniques, please read a book on molecular biology.

Applications of molecular epidemiology methods have completely changed the knowledge about infectious disease transmission for many microorganisms. The main application is within outbreak investigations. Being able to characterize the nucleic acid of the microorganisms permits an understanding of how the different cases relate to each other.

Molecular epidemiology methods have clarified the controversy about the origin of tuberculosis cases: Is it an endogenous (reactivation) or exogenous (re-infection) origin? On the one hand, endogenous origin postulates that *Mycobacterium tuberculosis* can remain alive in the human host for a lifetime and can start multiplying and producing lesions. On the other hand, exogenous origin theory postulates that re-infection plays a role in the development of tuberculosis. The immunity provided by the initial infection is not strong enough to prevent another exposure to *Mycobacterium tuberculosis*, and a new infection leads to disease. In countries with low tuberculosis transmission, for example, the Netherlands, most strains have unique RFLP fingerprints. Each infection is unique, and there are hardly any clusters of infections resulting from a common source. Most cases are the result of reactivation. This is in contrast with areas of high endemicity where long chains of transmission can be identified with few RFLP fingerprinting patterns (Alland et al. 1994). In some areas, up to 50% of tuberculosis cases are the result of re-infection.

Numerous new immunoassays have been developed. They depend on an antigen-antibody reaction, either using a test antibody to detect an antigen in the patient's specimen or using a test antigen to detect an antibody in the patient's specimen.

An indicator system is used to show that the reaction has taken place and to quantify the amount of patient antigen or antibody. The indicator can be a radioactive molecule (radioimmunoassay [RIA]), a fluorescent molecule (fluorescent immunoassay [FIA]), a molecule with an attached enzyme that catalyzes a color reaction (enzyme-linked immunoassay [ELISA or EIA]), or a particle coated with antigen or antibody that produces an agglutination (latex particle agglutination [LA]).

The reaction can be a simple antigen/antibody reaction or a "sandwich" immunoassay where the antigen is "captured" and a second "read out" antibody attaches to the captured antigen. The antibody used may be polyclonal (i.e., a mixing of immunoglobulin molecules secreted against a specific antigen, each recognizing a different epitope) or monoclonal (i.e., immunoglobulin molecules of single-epitope specificity that are secreted by a clone of B cells). It may be directed against an antigen on an epitope (i.e., a particular site within a macromolecule to which a specific antibody binds).

51.2.2 Mapping as an Epidemiological Tool

Plotting diseases on a map is one of the very basic methods epidemiologists do routinely. As early as 1854, John Snow, suspecting water as a cause of a cholera outbreak, plotted the cases of cholera in the districts of Golden Square, St. James, and Berwick in London. The cases seemed to be centered around the Broad Street pump and less dense around other pumps. The map supplemented by other observations led to the experiment of removing the handle on the Broad Street pump and subsequent confirmation of his hypothesis (Frost 1936).

Geographic Information Systems (GIS) have been a very useful tool in infectious disease research. GIS are software programs allowing for integration of a data bank with spatial information. The mapping component includes physical layout of the land, towns, buildings, roads, administrative boundaries, zip codes, etc. Data may be linked to specific locations in the physical maps or to specific aggregates. A GIS system includes tools for spatial analysis. Climate, vegetation, and other data may be obtained through remote sensing and combined with epidemiological data to predict vector occurrence.

However, these tools should be used with caution. They can be useful to generate hypotheses and identify possible associations between risk of disease and environmental exposures. Because of potential bias, mapping should never be considered as more than an initial step in the investigation of an association. "The bright color palettes tend to silence a statistical conscience about fortuitous differences in the raw data" (Boelaert et al. 1998). See chapter ▶Geographical Epidemiology of this handbook.

51.2.3 Computer Reporting and Software Progress

Web-based reporting, use of computer programs, and developments of sophisticated reporting and analytical software have revolutionized epidemiological data collection and analysis. These tools have provided the ability to collect large amounts of data and handle large databases. However, this has not been without risks. It remains crucial to understand the intricacies of data collected to avoid misinterpretation. For example, one should be aware that diseases and syndromes are initially coded by a person who may not be very software proficient, using shortcuts and otherwise could enter data of poor quality.

51.3 Basic Concepts

Too often one sees epidemiologists and statisticians preparing questionnaires, carrying out surveys, gathering surveillance information, processing data, and producing reports, tables, charts, and graphs in a routine fashion. Epidemiology describes the distribution of health outcomes and determinants for a purpose. It is important to question the goals and objectives of all epidemiological activities and tailor these activities to meet these objectives.

The description of disease patterns includes analysis of demographic, geographical, social, seasonal, and other risk factors.

Age groups to be used differ depending on the disease; for example, diseases affecting young children should have numerous age groups among children; sexually transmitted diseases require detailed age groups in late adolescence and early adulthood. Younger age groups may be lumped together for diseases affecting mainly the elderly. Gender categorization, while important for sexually transmitted diseases and other diseases with a large gender gap (such as tuberculosis), may not be important for numerous other diseases.

Geographical distribution is important to describe diseases linked to environmental conditions but may not be so useful for other diseases.

51.3.1 Biology or Natural History of Infectious Diseases: The Intersection of Biology, Microbiology, Climatology, Ecology, and Epidemiology

The natural history of an infectious disease is the way in which the disease is transmitted, how it develops over time from the earliest stage of its prepathogenesis phase to its termination as recovery, disability or death in the human population, in the absence of treatment or prevention.

Epidemiologists dealing with an infectious disease issue are best served by taking the time to study the natural history or biology of that specific infectious disease. Facts to be studied are the nature of the infectious agent (parasite, bacteria, fungus, virus, or prion), the natural hosts, mode of entry into the host and exit from the host, distribution in the host tissues, incubation period, signs and symptoms of illness, natural reservoir in animals or environment, resistance to environmental factors, and geographical distribution of the agent and of human illness (which may be slightly different).

51.3.2 Infectious, Communicable, Contagious, Transmissible Diseases

An infectious disease is a disease due to a specific infectious agent or its toxic products that arises through transmission of that agent or its products from an infected person, animal, or reservoir to a susceptible host, either directly or indirectly through an intermediate plant or animal host, vector, or the inanimate environment (Porta et al. 2008).

Infectious diseases are caused by an infectious agent (helminth, protozoa, fungus, bacteria, virus, or prion, sometimes referred to as microorganisms, although helminths are really not microorganisms).

This definition is apparently simple but may get more complicated:

- The infectious agent does not need to be present all the time. The infectious agent may trigger a pathological process that will continue on its own, even after the agent is gone.
- Other factors may be necessary to trigger the disease; the infectious agent alone cannot cause the disease. The infectious agent may be necessary but not sufficient for the infectious disease. Most agents causing opportunistic infections in AIDS patients cannot cause any disease in normal individuals. They can only cause disease if the host is severely immune-compromised.

The term *communicable disease* is specific to those diseases that can be transmitted from an infected individual to another one directly or indirectly. It is sometimes used interchangeably for infectious diseases (Porta et al. 2008). Sometimes communicable diseases are defined as a subset of infectious diseases that can spread from person to person.

The term *transmissible or contagious disease* is often synonymous to communicable disease.

51.3.3 From Exposure to Disease

The infectious process may be broken down into the following steps. If the infectious disease agent does not gain a foothold, the person was only *exposed* and the infectious disease process ends. If the disease agent gains a foothold but no reaction is occurring, the person will be *colonized* but not infected. An infection occurs when the disease agent attaches itself to the epithelium and begins to multiply. The infectious disease agent will release cytotoxins which will damage the cells and injure the tissue which leads to the dissemination through the human body. Even after dissemination, humans might not show any signs and symptoms and are therefore considered *asymptomatic* or show clinical signs and symptoms and are then considered *symptomatic*.

Exposed means that a person is placed in a situation where effective transmission of an infectious agent could occur. Being exposed does not always mean that transmission did occur. For example, being in the same room as an infectious tuberculous patient is being exposed since tuberculosis is transmitted by droplet nuclei. However, being in the same room with a person with HIV does not meet the criteria for exposure because conditions are not met for transmission to occur.

Exposure definition relies on information that may not be all known:

- Being in the same room with a tuberculous patient means being exposed if the patient is infectious (pulmonary tuberculosis with positive sputum). If the patient is not infectious, then exposure does not occur.
- Sharing a meal that resulted in a food poisoning outbreak is being exposed. If we know that only the potato salad was contaminated, then only those who ate the potato salad were exposed.

Infection: The entry and development or multiplication of an infectious agent in the body of humans or animals. Infection is not synonymous with disease. Disease implies some signs and symptoms or some negative impact on the health status of the individual.

Colonization: Porta et al. (2008) define in the Dictionary of Epidemiology infection and colonization as the same concept. However, in hospital-acquired infection control programs (often abbreviated as "infection control"), a distinction is made between colonization and infection: Colonization is the presence of a microorganism in or on a host with growth and multiplication but *without* any overt clinical expression or immune reaction in the host at the time the microorganism is collected (Brachman 1998). In contrast, infection entails some reaction from the host, either only on an immunological level or on an immunological and clinical level.

A *carrier* is an individual that harbors a specific microorganism in the absence of discernible clinical disease and serves as a potential source of infection. A carrier may be an individual who is colonized, incubating the disease, infected and asymptomatic, or convalescent from acute disease. The period of the carrier status may be short or lengthy. The portal of exit may be urine, genital secretions, feces, and respiratory, or the carrier may not excrete the agent (agent is circulating in the bloodstream).

Clinical infection: Clinical infection may result in signs and symptoms. Some of these may be less obvious or very minor. At the end of the spectrum is the individual with no sign, no symptoms who has an *asymptomatic infection or subclinical infection*. Asymptomatic infection does not mean that "all is quiet." It may cover some very active processes as in the asymptomatic phase of HIV infection, tuberculosis infection, or Hepatitis B carrier state.

51.3.4 Case, Index Case, Primary Case, and Secondary Case

A *case* is an operational definition. It denotes usually a person with a specific infectious disease.

A *surveillance case definition* is not a clinical diagnosis, both have very different purposes. A surveillance case definition is usually very precise and fairly restrictive so as to eliminate subjectivity, as much as possible. It uses a fixed set of indicators to classify disease status regardless of differences between individuals. In contrast to that, a clinical diagnosis' purpose is to ensure best treatment options to the patient and the diagnostic procedures may therefore vary between individuals.

A *diagnosis* is an expression of the clinical judgment of the physician that leads to the therapeutic decisions to be taken.

An *index case* is the earliest documented case of a disease that is included in an epidemiological study or the very first case of an infectious disease that was identified in an outbreak.

A *primary case* is the first individual (case) who brought the infection in the group of population studied. The primary case is not always the index case. The index case may have triggered an investigation, and in the course of the investigation, the primary (or original) case is identified.

A *secondary case* is a case that was infected from the primary case and consequently occurred at a later date. There may be tertiary cases and so on. Usually one does not define cases further than secondary cases. If cases are somewhat synchronized, one may speak of *generations* or waves of cases.

51.3.5 Source, Reservoir, Vehicle, and Vector

A *reservoir* is any person, animal, plant, or environmental medium (soil, water) in which the microorganism normally lives and multiplies, on which it depends primarily for survival, and where it reproduces itself in such a manner that it can be transmitted to the susceptible host. Consider the following examples: Humans are the only reservoir for *Mycobacterium tuberculosis*, measles, chickenpox and smallpox. Numerous animal species are reservoirs for *Salmonella*; rodents are reservoirs for plague. Surface water and water systems are reservoirs for *Legionella*. Soils and the gut of some animals (horses) are reservoirs for tetanus bacteria (*Clostridium tetani*).

A *source of infection* is the actual person, animal, or object from which the infection was acquired.

A *source of contamination* is the person, animal, or object from which environmental media are contaminated. For example, the cook is the source of contamination of the potato salad.

A *vehicle* is an inanimate object which serves to communicate disease, for example, a glass of water containing microbes or a dirty rag.

A *vector* is a live organism that serves to communicate disease. Best known examples are *Anopheles* mosquitoes and malaria as well as *Ixodes* ticks and Lyme disease.

51.3.6 Transmission and Chain of Infection

When describing transmission, one should consider the source of the infectious agent and the portal of entry in the human.

51.3.6.1 The Source of Infectious Material

There are very different sources from where the potential infectious material is coming from. It might be blood splashed on a medical employee during a procedure or a person coming in contact with someone else's blood after a motor vehicle accident. It might be internal body fluids (such as cerebrospinal, pericardial, pleural, peritoneal, synovial, and amniotic fluids), and most of these exposures would occur in the medical setting. For genital fluids (vaginal, prostatic secretions, semen), sexual contact is the main mode of transmission through mucous membranes. Furthermore, transmission of Hepatitis B virus (HBV) and herpes simplex virus (HSV) to the newborn can occur during delivery as the newborns are exposed to vaginal secretions. Both internal and genital fluids can contain blood-borne pathogens (such as HIV, Hepatitis B virus, Hepatitis C virus (HCV), and cytomegalovirus (CMV)). Both secretions (saliva, nasal discharge, sweat, tears, breast milk) and excretions can be infectious. Urine might be contaminated with schistosoma eggs or leptospira bacteria, and feces can contain numerous enteropathogens. Persons can be infected via sexual contact of mucosal membranes (nasal, oropharyngeal, rectal, genital). Contact with contaminated tissue can occur in transfer of human or animal tissue: blood transfusion, blood components (factor VIII), organ transplants, or tissue grafts. Some hormones and proteins may be extracted from the tissue but still carry the infectious microorganisms, for instance, prions of Creutzfeldt-Jakob disease (CJD) in human growth hormone extract. The rabies virus is normally transmitted through animal bites, but also human bites could potentially (however never documented) infect the bite victim with Hepatitis B or C virus. Last but not least, environmental materials such as food, water, air, or even contaminated dust play a major role in the transmission of infectious diseases.

51.3.6.2 The Portal of Entry into Humans

Infectious disease agents can enter the human body through very different paths. They can be inhaled with the air (the respiratory system). Eating contaminated food and drinking contaminated water (gastrointestinal system) can infect persons and of course through sexual activities. Transplacental or intrauterine transmission will pose a risk for the fetuses. Persons also can be infected with viruses, bacteria, rickettsia, and parasites through arthropod bites such as mosquito or tick bites.

51.3.7 Classification of Transmission

51.3.7.1 Droplet Transmission

There are many infectious diseases which are transmitted by droplets (see Box 51.1). Droplets are generated in the upper respiratory tract during talking, singing, spitting, sneezing, and coughing. They are also produced during suctioning, sputum induction, bronchoscopy, and other respiratory procedures. The droplets produced vary in size from 1 to 100 micron (μm). Droplets will fall to the floor; the speed of fall is related to droplet size (see Table 51.2).

Box 51.1. Infections transmitted by droplets

Haemophilus influenzae, *Neisseria meningitidis*, *Streptococcus pneumoniae*, *Corynebacterium diphtheriae*, *Yersinia pestis* (pneumonic plague), *Bordetella pertussis*, *Mycoplasma pneumoniae*, *Streptococcus group* A (pharyngitis, pneumonia, scarlet fever), adenovirus, influenza, mumps virus, parvovirus 19, and rubella virus

Table 51.2 Droplet falling rates

A droplet of (μm)	Will fall in	
100	10 seconds	Droplets above 5 μm are trapped in the nose and upper respiratory tract and usually do not make it to the bronchi
40	1 minute	
20	4 min	
10	20 min	
5–10	30–45 min	May reach the lower respiratory tract
≤5	Droplet nuclei	May be inhaled into the alveoli

Droplet transmission occurs by direct hit when these droplets are propelled from the infected host to the recipient's mouth, nasal mucosa, or conjunctivae. As a rule of thumb, this method of transmission is common within 3 feet of the infected patient. Inhalation of a droplet occurs also while it floats; however, this occurs only during a short period of time since the droplet is falling to the floor. Contact with surfaces contaminated with droplets is the main mode of transmission for rhinoviruses and respiratory syncytial viruses (RSV). Concentrations of rhinoviruses are much higher on the hands than in aerosols. Droplets are created by aerosolization of infectious material: The success of such aerosols reaching susceptible individuals depends on the environmental conditions: humidity, temperature of the air, air currents, and distances of the host. The use of suction devices, catheters in intensive care units (ICU), and blood products in hemodialysis may produce some aerosols containing infectious particles.

In nature, soil particles contaminated with rodent urine have been aerosolized and thought to be responsible for the transmission of hantaviruses. *Legionella* are frequently present in waters (surface waters, hot water systems, and condensation from air conditioning or ventilation systems). When water is sprayed (cooling towers, showers, and cool mist over produce), aerosols containing *Legionella* are generated.

The degree of infectivity depends on the microorganism concentration in the droplets emitted. This varies from one virus to another, from one strain to another. The infecting dose is variable. For some viruses, it may be quite small: seven virions for adenoviruses. Experiments made with influenza virus showed that for similar viral titer in lung tissue, some strains will have very high titer in bronchial secretions while others will not (Schulman 1970).

51.3.7.2 Airborne Transmission

Droplet nuclei or dust particles are responsible for this mode of transmission. Droplet nuclei are small droplets less than 5 μm in diameter. They result from evaporation of larger droplets or from direct formation of smaller droplets (particularly during coughing or during aerosol generating medical procedures). The transmission may occur over a long distance from the source patient.

Tuberculosis (TB) is one of the most important diseases transmitted by airborne means (see Box 51.2). Active pulmonary tuberculosis cases with acid-fast bacilli (AFB) on sputum smear are the cases that are infectious. Tuberculosis is almost exclusively transmitted by droplet nuclei (small particle of 1–5 μm) that contain *Mycobacterium tuberculosis*. The droplet nuclei must reach the pulmonary alveoli to start an infection. Large droplets are swallowed or get stuck in the trachea and bronchus; from there they are brushed back up and swallowed. The rare TB bacilli reaching the stomach are inactivated there. The role of droplet nuclei in the transmission of tuberculosis was demonstrated in several studies. In 1956, Riley and colleagues showed that air coming from rooms occupied by TB patients could infect guinea pigs (Riley et al. 1956). Coughing is the major producer of droplet nuclei. Speaking and singing also produce droplet nuclei, but these do not last very long (see Table 51.3).

Pulmonary tuberculosis cases may have up to 10,000,000 TB bacilli/milliliter (ml) of sputum. A typical sputum smear is about a hundredth of one ml (0.01 ml) and is covering about a 10,000 high power field (magnification ×100 for the oil immersion lens and ×10 for the eye piece). The probability of finding an acid-fast bacillus depends on the concentration of AFB in the sputum, and the number of microscopy fields examined (Toman 2004) (see Table 51.4). In a study carried out among

Box 51.2. Infections transmitted by droplet nuclei

- Tuberculosis (*Mycobacterium tuberculosis*)
- Measles (*Morbilli virus*)
- Chickenpox and shingles (*Varicella zoster* including disseminated zoster)

Table 51.3 Droplet production with coughing and singing

One good cough	⇒ 465 droplet nuclei
30 min after	228 are still airborne (49%)
Counting from 1 to 100	⇒ 1,764 droplet nuclei
30 min later	106 are still airborne (6%)

Table 51.4 Number of AFB per smear and number of immersion fields to be screened

No. of bacilli/ml/sputum	No. of AFB/smear	No. of immersion fields/AFB
10,000	100	100
100,000	1,000	10
1,000,000	10,000	1

Table 51.5 Proportion of infected contacts after index case coughed

Index cases coughed	Proportion of infected contacts (%)
>48 coughs/night	44
<12 coughs/night	27

Box 51.3. Infections transmitted by contact

- Gastrointestinal, respiratory, skin, wound infections
- Colonization with multidrug-resistant bacteria
- Enteric infections, enteroviral infections in infants
- Respiratory syncytial virus (RSV), parainfluenza
- Infectious skin infections: herpes simplex virus (HSV), impetigo, cellulitis, scabies, staphylococcal furunculosis
- Viral hemorrhagic conjunctivitis, viral fevers
- Some respiratory infections, bronchiolitis in infants, children
- Abscess, draining wound

contacts of smear-positive pulmonary cases, Loudon and colleagues showed that the more the index case coughs, the more infected individuals are to be observed among the close contacts (Loudon et al. 1969) (see Table 51.5).

51.3.7.3 Direct and Indirect Contact Transmission

Direct contact transmission results from a direct body surface to body surface contact and physical transfer of microorganisms. Direct contact occurs when shaking hands, taking pulse, turning a patient over, and having sexual intercourse.

Different viruses and bacteria can be transmitted by contact (see Box 51.3). Indirect contact transmission involves contact with the intermediate of an object. Indirect contact occurs through a contaminated dressing, instrument, or glove as well as door handles and keyboards.

51.3.7.4 Gastrointestinal Transmission: Fecal-Oral Route

Transmission by the fecal-oral route is the second most important mode of transmission after the respiratory tract for several infectious disease agents (see Box 51.4). The fecal-oral route refers to the mode of transmission of microorganisms excreted

Box 51.4. Infections transmitted by gastrointestinal transmission: fecal-oral route

- Typhoid fever
- *Shigella* spp.
- Cholera (*Vibrio cholerae*)
- Polio
- Coxsackie virus, echovirus, reovirus
- Norovirus
- Rotavirus
- Hepatitis A, Hepatitis E

by the feces and transmitted to the oral portal of entry through contaminated food, water, milk, drinks, hands, and flies.

The site of entry may be the oropharynx for some microorganisms or the intestinal tract for most viruses. *Surviving through the upper GI tract is essential*. Viruses with envelopes do not survive exposure to hydrochloric acid in the stomach, bile acids in the duodenum, salts and enzymes of the gut. Small enteroviruses without envelope (norovirus, rotavirus, polio, and coxsackie viruses) are able to resist. Hepatitis A and E are also transmitted by the fecal-oral route. For adenoviruses and reoviruses, this route is of minor importance.

Some of these pathogens are essentially found in humans (*Shigella*), while others may survive or multiply in the environment for long periods of time (*Vibrio cholerae*, poliomyelitis virus). This mode of transmission is more amenable to control measures than the respiratory route. Good personal hygiene (mostly proper hand washing), purification of drinking water, pasteurization of milk and dairy products, and sanitary preparation of food are all highly effective prevention measurements for these types of infectious diseases.

51.3.7.5 Gastrointestinal Transmission: Animal Host and Contaminated Food Products

Salmonellas infect a wide variety of domestic animals, birds, and other wildlife. Foods derived from salmonella-infected animal (eggs, dairy products, meat) are the major source of infection if improperly prepared. Salmonella is less often transmitted by water or direct contact. Other microorganisms such as Campylobacter, Yersinia, and Listeria are also transmitted through contaminated food products (see Box 51.5).

Food poisoning overlaps both classes of gastrointestinal transmission. Food poisoning may result:

- From consumption of food from an infected animal or undercooked eggs, for example, chicken and eggs with *Salmonella* or *Listeria* in unpasteurized milk

Box 51.5. Infections transmitted by gastrointestinal (GI) transmission: animal host and contaminated food products

- Salmonella
- Campylobacter
- Yersinia
- Listeria

- From consumption of food contaminated in the environment, for example, *Vibrio vulnificus* or *Vibrio cholerae* in raw oysters or undercooked seafood
- From food contaminated during preparation from an infected food item, for example, potato salad contaminated with *Salmonella* from raw chicken because the uncooked chicken and the salad ingredients were cut on the same cutting board
- From food contaminated by a human source, for example, typhoid fever carrier

51.3.7.6 Skin or Mucous Membrane Transmission

Transmission through the skin is the third most common mode of transmission of infection. Penetration through the intact skin is unlikely. Break in the skin barrier may result from needle injection, cut during a surgical procedure, accidental cut, crushing injury, and bite (rabies).

Transmission of blood-borne pathogens (Hepatitis B and C viruses (HBV, HCV) and HIV) does not occur if the blood was splashed exclusively on intact skin. Penetration through the skin is necessary. In the case of HIV, it takes injury with a hollow bore needle or other sharp object (lancet, glass, and scalpel) with blood to cause an infection. Solid needles do not carry sufficient quantities of blood to cause an infection. The viral titer is the best predictor of risk of infection. After percutaneous exposure to blood from infected patients, the risk of infection in the recipient is 30% for HBV (eAntigen positive), 3% for HCV, and 0.3% for HIV. This follows the ranking of viral titers.

Mucosal membranes allow penetration by blood-borne pathogens. Data from 21 studies worldwide on mucosal membrane exposure to HIV showed only one conversion in a total of 1,107 health-care workers (HCWs). The proportion of conversion was 0.09% (1/1,107).

Some parasites are able to penetrate actively through the intact skin: hookworm larvae and schistosoma cercariae.

51.3.7.7 Sexual Transmission (Mucous Membrane Transmission)

The genital tract is a special case for transmission through the mucosal membranes. The bacteria and viruses listed are present in the genital fluids and on the mucosal membranes (see Box 51.6). They may be transmitted to the mucosal membranes of the partner during sexual acts: Membranes involved may be the vagina, penile urethra, anus and rectum, or oropharynx. Some of microorganisms such as *Shigella* spp. and *Campylobacter* spp. are primarily considered to be transmitted to the gastrointestinal (GI) tract. However, due to transmission when the rectum is involved in sexual activities, they are also listed as sexually transmitted disease (STD) agents.

The presence of lesions on the recipient partner seems to predispose to acquisition of infection, particularly for HIV.

51.3.7.8 Perinatal Transmission (Mucous Membrane Transmission)

These infections (see Box 51.7) occur when the newborn goes through the birth canal, from the cervix or vagina to the newborn.

Box 51.6. Sexual transmissions (mucous membrane transmission)

- *Neisseria gonorrhoeae*, *Chlamydia trachomatis*
- *Treponema pallidum* (syphilis)
- *Hemophilus ducreyi*
- *Mycoplasma hominis*, *Ureaplasma urealyticum*
- *Calymmatobacterium granulomatis*
- *Shigella* spp., *Campylobacter* spp.
- Group B streptococci
- Bacterial vaginosis-associated bacteria
- Herpes simplex virus (HSV) 1 and 2
- Cytomegalovirus (CMV) or herpes virus 5
- Hepatitis B virus (HBV)
- Human papilloma virus
- Molluscum contagiosum virus
- HIV (human immunodeficiency virus) 1 and 2
- *Trichomonas vaginalis*
- *Entamoeba histolytica*, *Giardia lamblia*
- *Phthirus pubis*
- *Sarcoptes scabiei*

Box 51.7. Perinatal transmission (mucous membrane transmission)

- *Neisseria gonorrhoeae*
- *Chlamydia trachomatis*
- HBV
- HSV

Box 51.8. Transplacental transmission or vertical transmission

- *Treponema pallidum* (syphilis)
- *Toxoplasma gondii*
- CMV, HBV
- HIV
- HSV
- Rubella, varicella

51.3.7.9 Transplacental Transmission or Vertical Transmission

The microorganisms in this case are present in the blood of the mother and are able to go through the placenta to infect the fetus (see Box 51.8). In some cases, it is difficult to differentiate between perinatal or transplacental transmission, since both modes of transmission are known to occur.

51.3.7.10 Urinary Transmission

Although some bacteria (typhoid fever, leptospirosis) and viruses (CMV, measles) may be excreted in the urine, the role of urine is a minor one in the transmission of diseases. In urinary schistosomiasis, the adult worms live in the venous plexus around the urinary bladder. They lay their eggs in the lining of the bladder. The eggs are excreted in the urine. If they reach water, they hatch into larvae which look for a suitable intermediate host (freshwater mollusk).

51.3.7.11 Arthropod-Borne Transmission

Mosquitoes, flies, fleas, true bugs, ticks, and lice may transmit various microorganisms by two mechanisms (see Table 51.6):

1. Passive transmission: the insect acts as a live syringe. It picks up microorganisms from blood or superficial lesions and passes them on to another human. There is no incubation time, no multiplication of microorganisms while carried by the arthropod. This mode of transmission is not specific; a wide variety of microorganisms may be transmitted, but the transmission is not very efficient.
2. Active transmission involves multiplication of the microorganisms in the arthropod. This applies only to some microorganisms with a definite set of arthropods. This mode of transmission may be very effective: The microorganisms may be multiplied a thousand to a million times. This mode requires a period of multiplication in the arthropod.

51.3.7.12 Common Vehicle Transmission

The microorganisms have contaminated the "common vehicle" and are persisting over a long period of time in/on the common vehicle. The common vehicle can

Table 51.6 Arthropod-borne diseases

Disease (infectious agent)	Vector/intermediate host
Bacteria	
Plague (*Yersinia pestis*)	Fleas
Borrelia	
Lyme disease (*Borrelia burgdorferi*)	*Ixodes* ticks
Relapsing fever (*Borrelia recurrentis*)	*Ornithodoros* ticks
Rickettsia	
Epidemic typhus (*Rickettsia prowazekii*)	Lice, Pediculus humanus
Murine typhus (*Rickettsia typhi*)	Fleas
Scrub typhus (*Rickettsia tsutsugamushi*)	Larval mites
Rickettsialpox (*Rickettsia akari*)	Mouse mite
Rocky Mountain spotted fever (*Rickettsia rickettsii*)	*Dermacentor, Amblyomma* ticks
Trench fever (*Rickettsia quintana*)	Lice
Virus	
Dengue	*Aedes aegypti*
Flaviviridae	Mosquitoes
St. Louis encephalitis	*Culex* mosquitoes
Japanese encephalitis	*Culex* mosquitoes
Tick-borne encephalitis	*Ixodes* ticks
Powassan	*Dermacentor* ticks
Togaviridae: Alphavirus	
Eastern equine encephalitis	Mosquitoes
Western equine encephalitis	Mosquitoes
Venezuelan equine encephalitis	Mosquitoes
Bunyaviridae	
California encephalitis, La Crosse	Mosquitoes
Crimean-Congo hemorrhagic fever	Ticks
Kyasanur forest hemorrhagic fever	Ticks
Yellow fever	*Aedes* mosquitoes
Protozoa	
Malaria (*Plasmodium* sp.)	*Anopheles* mosquitoes
Chagas disease (*Trypanosoma cruzi*)	*Triatoma* sp. (bugs)
Sleeping sickness (*Trypanosoma gambiense*)	*Glossina* sp. (tsetse flies)
Leishmaniasis (*Leishmania* sp.)	*Phlebotomus* (sandflies)
Helminths	
Bancroft's filariasis (*Wuchereria bancrofti*)	*Culex*, *Aedes*, and *Anopheles* mosquitoes
Malayan filariasis (*Brugia malayi*)	*Culex*, *Aedes*, and *Anopheles* mosquitoes
Mansonella ozzardi	*Culicoides*, *Simulium*
Acanthocheilonema perstans	Gnats, *Culicoides*
Onchocerciasis (*Onchocerca volvulus*)	Black gnats, *Simulium*
Loiasis (*Loa loa*)	Mango flies, *Chrysops*

be food, water (either drinking the water or swimming in the water), soil (tetanus bacteria), medications, medical devices, or equipment.

51.3.8 Different Roles in Transmission: Indicator, Maintenance, and Amplifier

For some infectious diseases, different segments of the population play different roles. The best example for that is foot-and-mouth disease (FMD), a viral disease which rarely affects humans. Cloven-hoofed animals (such as cattle, goats, sheep, and pigs) are susceptible to FMD. FMD viruses are transmitted by air from one infected animal to another. Pigs are considered to be the amplifier host because they may exhale up to 1 million viral particles/ml of air. Sheep are an important reservoir of the virus and are considered maintenance hosts. They are usually asymptomatic when infected with foot-and-mouth disease. When these sheep mix with cattle, the cattle develop severe clinical signs and are therefore easily detected (*indicator host*).

51.3.9 Incubation Period, Latent Period, and Serial Interval

51.3.9.1 Incubation Period

The incubation period is the time interval between the invasion by a microorganism and the first signs or symptoms of disease (onset of disease). The concept of incubation period relies on the assumption that the disease is not asymptomatic and that the onset is clearly identifiable. For asymptomatic cases or carriers, incubation periods are irrelevant. For some infections, a person may get exposed to the agent, become colonized, and sometime in the future become a case. If this happens, incubation is also irrelevant. Incubation periods are only useful if infection is followed by disease within a certain period of time. Incubation periods are useful tools when carrying out infectious disease investigations. A person usually can tell when the first symptoms of a specific disease appeared. From that date, subtracting the incubation period, epidemiologists may estimate the date of infection (within a certain interval). It is also important for follow-up on potential contacts to the primary case that the primary case might have been already infectious before exhibiting any clinical signs and symptoms (see Fig. 51.1). In many instances, a person may be infectious toward the end of the incubation period but before the appearance of the first symptoms. The incubation period varies according to numerous factors:

- Portal of entry: The closer the portal of entry to the site of disease, the shorter the incubation period.
- Type of infection (local or systemic): Diseases caused by local multiplication of a microorganism have short incubation periods. Those that require systemic dissemination and secondary localization have longer incubation periods.
- Pathogenesis: Diseases due to a preformed toxin have very short incubation periods. Diseases due to direct involvement of epithelial surfaces have short incubation periods, for example, streptococcal sore throat, bacterial pneumonias,

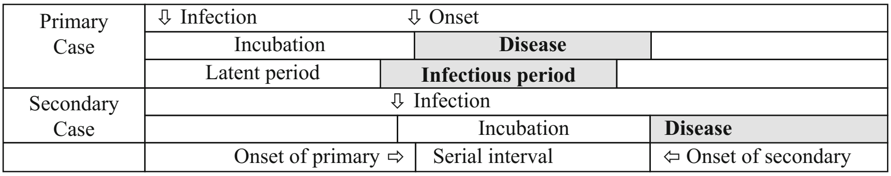


Fig. 51.1 Incubation, latent period, and serial interval

shigellosis, cholera, and gonorrhea. In contrast, *Mycoplasma pneumoniae*, diphtheria, and pertussis as well as diseases like syphilis, brucellosis, and typhoid fever have long incubation periods (2–3 weeks).

- Immune status of the host: It is important that the notion of incubation is relative. HIV provides a good example. Infection of an individual with HIV is followed by a flu-like syndrome. It includes fever, headache, miscellaneous aches (neck and back), malaise, lymphadenopathy, and rash. The incubation period for this primary syndrome is 2 weeks to 2 months. The patient then enters into a remission period with no clinical signs. However, during this period, the HIV multiplies at variable rates, destroying CD4+ lymphocytes which are generated as fast as they are destroyed. The latent remission ends when the patient's organism is no longer able to produce CD4+ lymphocytes in sufficient quantities. Immune defenses fail rapidly and opportunistic infections develop. This phase is considered as the AIDS (acquired immune deficiency syndrome) disease. The incubation period for AIDS diseases ranges from 2 to 10 years, with less than 10% having an incubation period greater than 10 years.

In rabies, the incubation period depends on the length of time it takes the virus to progress along the neurons to reach the brain. Once the brain is reached, the disease becomes manifest. The incubation may be as short as 9 days if the bite was in the face or as long as 1 or 2 months if the bite occurred in the leg. The longest incubation period known for rabies virus is 9 years.

The incubation period is useful for tracing the source of infection and contact, determine the period of surveillance, allow for prophylaxis to become effective (diseases with a long incubation period may be prevented by immunization if administered early), identification of point source or propagated epidemics.

Incubation period in a vector is the time interval between entry of the microorganism in the vector and the time the vector becomes infective. This is also called the extrinsic incubation in contrast to the intrinsic incubation period in humans.

51.3.9.2 Latent Period

The latent period of infection is the length of time between infection and the beginning of the infectious period. It is also a period during which no symptoms occur, an asymptomatic window in the disease (latent period of syphilis, of HIV infection).

51.3.9.3 Serial Interval

A serial interval for diseases spread from person to person is the time between successive generations of cases, that is, the time between appearances of symptoms in successive generations. If a person is infectious before onset of symptoms, the serial interval may be lower than the incubation period.

51.3.9.4 Infectious (Infectivity) or Communicability Period

The infectious (infectivity) period is the length of time a person may transmit a microorganism. There are several patterns for infectious periods:

- Short period at the end of the incubation period and at the beginning of the disease (measles, chickenpox)
- Short period and a few individuals become chronic carriers (Hepatitis B)
- Throughout the disease (open cases of active pulmonary tuberculosis, malaria).

Measuring infectivity is difficult. It is seldom the result of well-controlled studies. It is often the interpretation of observational studies on the occurrence of secondary cases. Factors such as amount of infectious agents put out by the source, closeness, length of contact, and susceptibility of the target contacts have to be considered. In recent times, nucleic acid testing has been used to find remnants of infectious disease agents in human or environmental materials, but their significance to transmission is difficult to interpret.

51.3.10 Distribution Pattern in the Population

For a better understanding of the distribution of infectious diseases in populations, the below terms have to be defined: Epidemiologists define *sporadic* cases as the occurrence of single illnesses in irregular or random instances. *Endemic* defines the occurrence of cases of an illness with a constant frequency. Depending on the intensity of the occurrence, the terms holoendemic, hyperendemic, or hypoendemic are used. *Epidemic is defined as the* occurrence in a community of cases of an illness with a frequency clearly in excess of normal expectancy. If this occurrence of an epidemic occurs worldwide or affects numerous countries, epidemiologists consider it a pandemic. The most recent pandemic was declared in June 2009, when the WHO declared a pandemic of novel influenza A (H1N1). At the time, more than 70 countries had reported cases of novel influenza A (H1N1) infection, and there were ongoing community level outbreaks of novel H1N1 in multiple parts of the world. An *outbreak* is defined as two or more related cases with the identical infectious disease agent suggesting the possibility of a common source or transmission between these cases. It also could be defined as a very limited epidemic; however, the word "epidemic" is usually avoided when the number of cases is relatively small so as not to scare the public. *Elimination of disease* is the reduction to zero of the disease incidence in a defined geographical area (e.g., neonatal tetanus) compared to the *elimination of infections* which is defined as the reduction to zero of incidence of infection in a defined geographical area (e.g., measles, poliomyelitis).

Table 51.7 Infectious dose and attack rates

Dose (no. of organisms)	Attack rate (%)
Experimental human salmonellosis	
125,000	17
695,000	33
1,700,000	67
(McCullough and Eisele 1951)	
Typhoid fever	
1,000	0
100,000	28
10,000,000	50
100,000,000	89
1,000,000,000	95
(Hornick et al. 1970)	

If there is a permanent reduction to zero in the worldwide incidence, the disease is considered *eradicated*, such as smallpox was in 1980.

51.3.11 Infectious Dose

The dose of pathogens received by the exposed individual is an important aspect of infectivity. There is also a close correlation between dose and type of contact. A closer, more direct type of contact delivers a higher dose.

It is rarely possible to have an exact measure of the infecting dose. In the past, experiments have been carried out with human volunteers. In one experiment, *Salmonella bareilly* was given to several groups of six volunteers (McCullough and Eisele 1951). A case was defined as one experiencing clinical diarrhea with *S. bareilly* isolated from the stools. Some of the cases excreted *Salmonella* for 1 day, some for 2 days. The corresponding attack rates, that is, percentage of volunteers experiencing a clinical diarrhea, are displayed in Table 51.7. The attack rate depends heavily on the working case definition.

From this type of data, one may calculate an infectious dose 50 (ID50) = the dose of pathogenic microorganism that will cause disease in 50% of the susceptible exposed. In some outbreaks, particularly foodborne outbreaks where contaminated food is saved, it may be possible to estimate the infectious dose. The dose may also be important in determining the severity of disease. To give an example, 1,000 *Vibrio cholerae* bacteria produce asymptomatic infections, 10,000 to 1 million bacteria produce simple diarrhea in 60%, and at least 1 million bacteria produce severe diarrhea with dehydration in 25–50% of volunteers.

51.3.12 Environmental Factors

There are several factors which influence the spread of microorganisms in the environment. The spread of infectious diseases depends on:

1. The stability of the microorganism in the physical environment required for its transmission including resistance to desiccation, high or low temperature, and ultraviolet light
2. The amount of microorganisms in the vehicle of transmission
3. The virulence and infectivity of the microorganisms
4. The availability of the proper vector or medium for the transmission

Environmental characteristics play a role on different levels:

1. Survival of the virus in the environment
2. Influence on the route of transmission
3. Influence on the behavior of the host

A warm environment enhances the transmission of microorganisms transmitted by water. In tropical and temperate areas, summer increases contacts between humans and surface water. Summer brings more people outside, particularly in the evening, and increases contacts between humans and mosquitoes and other arthropod vectors.

In the cooler seasons in temperate climates, in the rainy season in tropical climates, people tend to stay and congregate indoors promoting transmission by airborne or droplet mechanisms. Long stays in the hot and dry environment indoors impair the protective mechanisms of human mucous membranes and may facilitate the attachment of viruses onto the upper respiratory mucous membranes. The incidence of upper respiratory infections is as high in the middle of winter in the temperate climates as in the middle of the monsoon or rainy season in the tropical climates.

51.3.13 Host Factors

51.3.13.1 Extrinsic Host Factors

Exposure to infectious disease agents depends on both intrinsic (internal) and extrinsic (external) host factors. Extrinsic host factors are the method of transmission of the microorganism as well as the host behavior. Exposure to microorganisms which are transmitted by droplet or airborne modes is very common. Anyone who is out in the public is likely to be exposed to these microorganisms. Microorganisms which are transmitted by vectors result usually from special occupations or special settings (hobbies or leisure activities). For example, persons who love to be outdoors (camping, hiking, or working on fields or in the forest) are more likely to be bitten by ticks or mosquitoes and therefore more likely to develop one of the zoonotic diseases which are transmitted by these arthropods. Exposure to sexually transmitted microorganisms depends entirely on the sexual activities, number of sex partners, and/or lifestyle of the hosts and the carriers of these diseases.

51.3.13.2 Intrinsic Host Factors

The transmission of infectious diseases is also regulated by intrinsic factors that influence the host response. It depends on how many microorganisms are transmitted (dose), how virulent the strain is, and how the microorganisms enter the human body. The person's age at time of infection is important, too. In general, the probability of clinical disease increases with age (e.g., polio, Hepatitis). Preexisting level of immunity to the disease, the nutritional status of the host, as well as any preexisting disease will influence a successful transmission as well. Individuals with impaired immune response (HIV, patients on immunosuppressive therapy for cancer or transplant) have a higher risk of developing severe disease. Also personal habits or lifestyle factors such as smoking, drinking alcohol, drug abuse, or exercise can influence the host response. Smoking depresses the ciliary function of the bronchial tree and increases susceptibility to infections (e.g., tuberculosis). Alcohol consumption increases the risk for chronic Hepatitis infections. Also psychological factors such as motivation and attitude toward disease can contribute to the transmission of infectious disease agents.

51.4 Occurrence of Infectious Diseases

Especially in outbreak situations, epidemiologists investigate the occurrence of disease by asking the following questions:

- When did the disease occur (time)?
- Where do the cases come from (place)?
- Who got infected with the disease (person)?

51.4.1 Time

51.4.1.1 Epidemic Curve

An epidemic curve is the standard graphic representation of cases occurring over time. It is a histogram with number of cases plotted along the vertical axis and time along the horizontal axis. The time unit may be in hours (rapid outbreak such as foodborne outbreaks due to a toxi-infection), days, or weeks. It is important to have a good time unit applied; if the time unit is too short or too long, one does not get a visual picture of the outbreak dynamics.

The examples in Figs. 51.2 and 51.3 show the epidemic of a Saint Louis encephalitis (SLE) outbreak that occurred in Louisiana in September to October 2001. Figure 51.2 uses days as a time unit, and Fig. 51.3 uses weeks. Figure 51.3 provides a better understanding of the outbreak.

An epidemic curve may provide some clues about the nature of the outbreak. A point source outbreak is relatively contracted in time, while a continuous source outbreak is more stretched out. In the beginning of an outbreak which is due to person-to-person transmission, one may see the successive generations.

51.4.1.2 Seasonal and Annual Variations

Seasonal variations are important for some infectious diseases, particularly those which are heavily influenced by the environment such as water, food, and arthropod

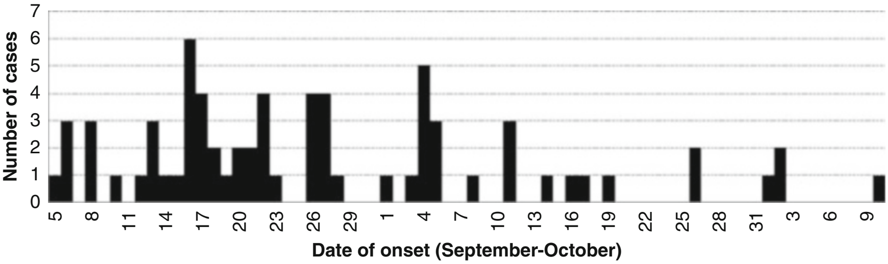


Fig. 51.2 Saint Louis encephalitis (SLE) outbreak, Monroe, Louisiana, September–October 2001, epidemic curve by day of onset

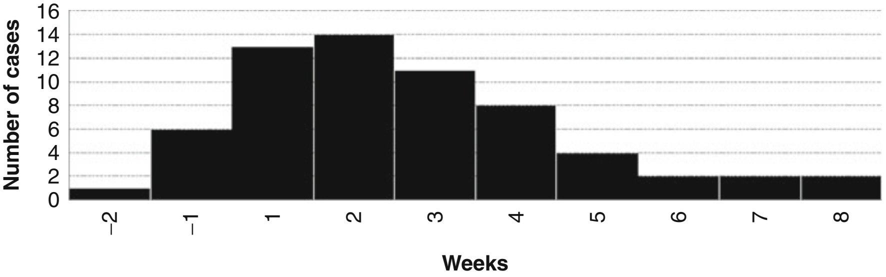


Fig. 51.3 Saint Louis encephalitis (SLE) outbreak, Monroe, Louisiana, September–October 2001, epidemic curve by week of onset. Week 1 = notification of SLE cases. Some cases actually had onset before notification (hence week −1 and week −2)

vectors. These are more prevalent during the warmer months of the year. Respiratory infections on the other hand are more prevalent during the winter in temperate areas or the rainy season in tropical areas.

Annual variations are thought to be mostly the result of accumulation of immune people after epidemics of an infectious disease. Once the proportion of the immune population has reached a certain threshold, there are very few susceptible individuals. In the absence of large epidemics, the pool of susceptible builds up back again, and herd immunity is down again. Then the circumstances are right for another epidemic. These cyclical patterns vary, every other year for measles before the advent of the vaccine, every 3–4 years for pertussis.

51.4.2 Place

Mapping cases is a very common tool used in infectious disease epidemiology. The map may range from a facility to a city, county, province, or country. Maps may be spot maps or rates in boundaries. Mapping may also provide some clues as to the etiology and evolution of an outbreak.

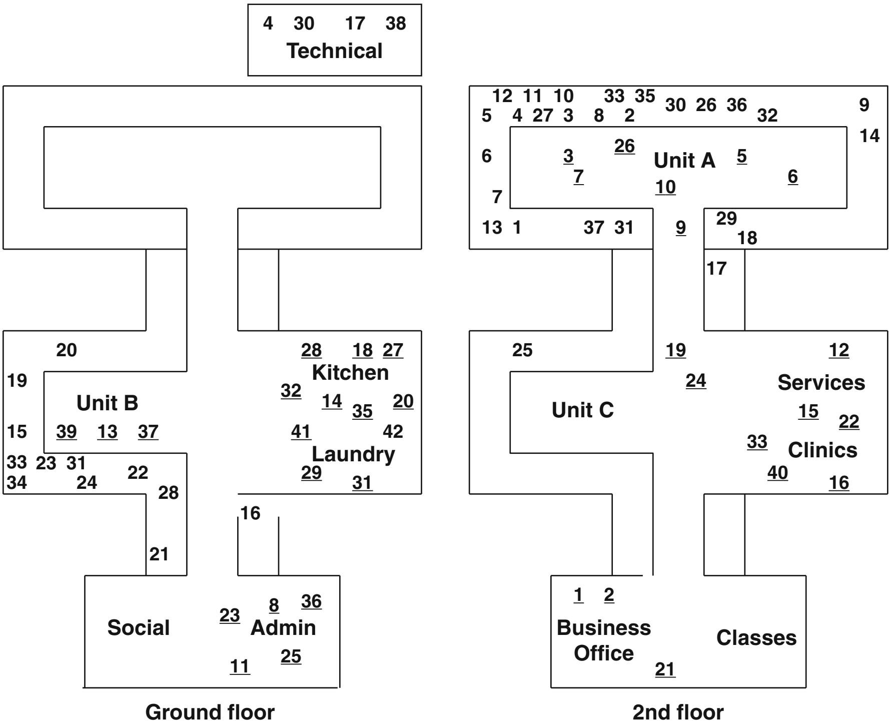


Fig. 51.4 Norovirus outbreak, nursing home, New Orleans 2005

Figure 51.4 shows a map of a nursing home outbreak that occurred in New Orleans in April 2005. Cases are plotted according to their location and are numbered in sequence as they were diagnosed. Cases among employees are underlined compared to cases among patients which are not underlined. The outbreak started with two unattached employees (as indicated as numbers 1 and 2 in Fig. 51.4); they infected an employee in A Unit (number 3). This was followed by a large outbreak in the Unit A. Then an employee in the Unit B was infected and soon after the outbreak was continuing in Unit A but also spreading in Unit B while it also spread among unattached employees. The Unit C was mostly vacant, so there were few cases.

The floor map is accompanied by an epidemic curve showing the cases by their location (A = patients in Unit A, a = employee in Unit A, similarly for unit B and C, and e for employee not attached to a specific unit) (Fig. 51.5).

51.4.3 Host Immunity

The immune status of individuals plays a major role in their susceptibility to infectious agents.

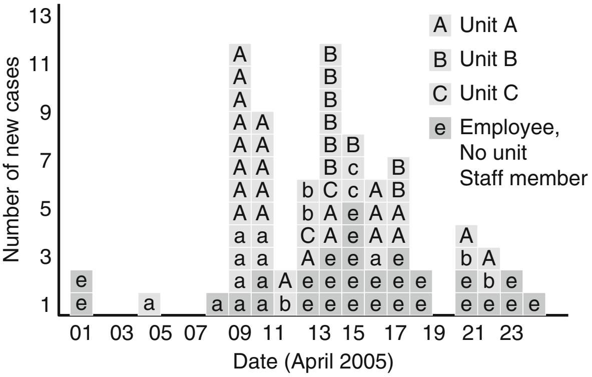


Fig. 51.5 Epidemic curve of a Norovirus outbreak in a nursing home, New Orleans, April 2005

51.4.3.1 Specific Immunity

After acquiring an infectious disease, the immune system reaction may (or may not) lead to protection against another attack of the disease. This so-called refractory period is the time period where no other infection of this disease can occur. It may last from a few days up to lifetime depending on the infectious agent. For some infections such as gonorrhea and *Chlamydia*, there is hardly any protection. A person may become re-infected soon after being cured of the previous infection. For some diseases such as syphilis, a person may not be super-infected while being already infected (immunity of premonition). Once treated, the immunity disappears. For others, the immunity may last for years or even a lifetime (measles, chickenpox). Lifetime immunity may be boosted by repeated contacts with the infectious agents. Immunity may have been acquired following an overt clinical bout of disease or following an unapparent infection. Eighty percent of children in the USA are immune to cytomegalovirus (CMV) infection, and the majority have had a completely asymptomatic infection.

The herd immunity is the immunity of the group. It is related to the sum of immune individuals over the total population. If a high proportion of a population is immune to a disease, one speaks of herd immunity. Above a certain threshold, the incidence of infections may decline. For example, invasive pneumococcal disease decreased dramatically after the pneumococcal conjugate vaccine (PCV7) for young children was introduced in 2000 in the USA. The modeled incidence of pneumococcal disease covered by this vaccine decreased by 76.6% in the unvaccinated population if the three-dose vaccination was completed in children before 15 months of age based on an estimated vaccine coverage between 38.1% and 54% in this population (Haber et al. 2007).

51.4.3.2 Immunocompetence

Immunocompetence is the ability of the immune system to respond to foreign substance and provide adequate protection. Humans with normally functioning immune systems are protected against a wide variety of infectious agents.

Immunocompetence is not fully developed in newborns and is weakened by age or by numerous chronic, acute diseases or medical treatments. Deep depression of the immune system is called immune deficiency. Infection by the HIV virus leads to a profound acquired immune-deficiency syndrome (AIDS). The spread of an infection may be very different depending on the prevalence and distribution of immune-deficient individuals. For example, levels of tuberculosis disease reach the highest incidence in countries with high prevalence of HIV infection.

51.4.4 Contacts: Patterns, Networks, and Structures

Contacts (the persons who are the recipients of the infectious agent) and contact patterns are important in infectious disease epidemiology. Contact may be defined as the type of interaction (or situation) between a person acting as a source of an infectious agent and a person susceptible when the interaction may lead to transmission of the infectious agent.

51.4.4.1 Contact and Interaction

The types of contact vary widely with the type of transmission. Direct contact occurs when the infected host and the susceptible recipient have their skin or mucous membranes touching, for example, by shaking hands, kissing, or having sexual intercourse. Indirect contact occurs when the transmission between both persons involve an inanimate object (fomite) or a mechanical vector (e.g., fly).

When exposure occurs through the air, droplets or droplet nuclei are the vehicles of the infectious agent. The circumstances of the "contact" require a precise definition. For example, a contact of an infectious pertussis case is a person who (1) had face-to-face interaction at less than 3 ft for at least 10–15 min or (2) shared confined space for 1 hour or (3) had a child in a crib located 3–6 ft away or (4) had direct contact with oral, nasal, or respiratory secretions or (5) had shared food, drink, or eating utensils or (6) kissed or (7) was in a medical setting during examination of mouth, throat, intubation, or cardiopulmonary resuscitation (CPR). This type of very detailed definitions is useful to determine the persons at risk of infection and place them under surveillance or prophylaxis.

51.4.4.2 Contact Patterns: Sociograms

Sociograms were developed to analyze choices or preferences within a group. They can diagram the structure and patterns of group interactions. A sociogram consists of nodes (people) and links (contact meeting the definition of a possible transmission).

The nodes on a sociogram who have many choices are called stars. Those with few or no choices are called isolates. Individuals who choose each other are known to have made a mutual choice. One-way choice refers to individuals who choose someone, but the choice is not reciprocated. Cliques are groups of three or more people within a larger group who all choose each other (mutual choice).

The following is an example of a sociogram for sexual contact patterns in a hypothetical high school (Fig. 51.6). Squares represent males, circles females, and

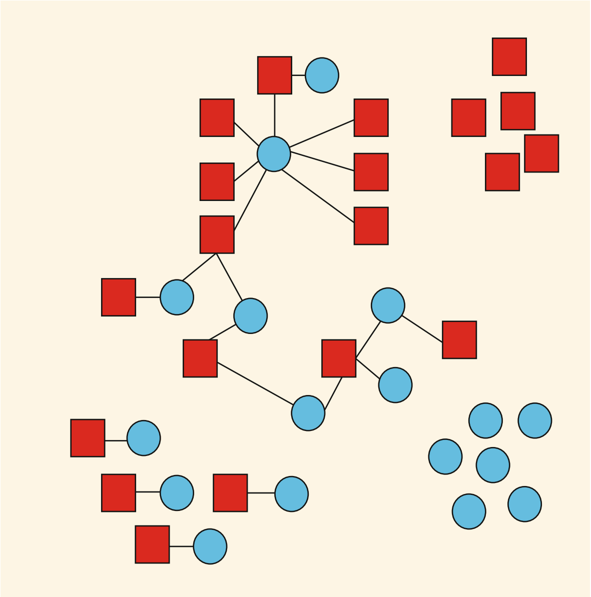


Fig. 51.6 Example of a sociogram depicting sexual relations in a hypothetical high school. Square = male, circle = female

links sexual relations. There are celibate males and females, isolated pairs and a large network of individuals having sexual relations, some with a single partner, and some with multiple partners. Such sociograms may be useful to describe and understand an outbreak, but it may also be useful to describe contact patterns in the absence of any specific disease. It would then help understand what would happen if an outbreak would occur in the population.

51.4.5 Risk Measures

Incidence rate (cumulative incidence), incidence density, and prevalence are commonly used (see chapter ▸Rates, Risks, Measures of Association and Impact of this handbook). The numerator may be the number of cases or the number of persons with serological evidence of past infections for example. Depending on the circumstances, the denominator may be the entire population or the number of persons exposed. All rates used in epidemiology are also used in epidemiology of infectious diseases. However, attack rate and case fatality rates are especially common to infectious disease epidemiology.

51.4.5.1 Attack Rate

The *attack rate* is the proportion of those exposed to microorganisms that develop the disease. Attack rates are frequently used in infectious disease epidemiology. They are heavily influenced on the used definition of exposure and disease. If a segment of the population is immune (previous natural infection or immunization), it will not be susceptible to the disease, and therefore, the attack rate will be underestimated. Attack rate is a misnomer. The attack rate is a cumulative incidence of cases that occurred during an outbreak.

51.4.5.2 Case Fatality Rate

The case fatality rate is the proportion of people who will die of a certain disease over those who have the disease. Since it is a rate, a time period has to be specified. It is different from the mortality rate which is the proportion of the entire population which dies from a certain disease during a definite period of time (usually 1 year).

51.4.5.3 Reproductive Rate

The reproductive rate is the average number of cases that will result from an index case. The reproductive rate depends on the:

- Probability of transmission in a contact between infected and susceptible
- Frequency of contacts in the population
- Duration of infection
- Proportion already immune in the population

51.4.6 Study Designs

Although any design is used in infectious disease epidemiology studies, the most common designs are descriptive and case-control studies (see previous sections in this handbook).

Case reports are detailed descriptions of single cases with exposure, clinical, treatment, and other relevant information. Description of single cases exposed under unusual circumstances may have profound consequences on the prevention of infectious diseases (case of HIV transmitted by a dentist, cases of rabies transmitted by unspecified contact with bats, case of West Nile virus infection transmitted by transfusion or transplantation).

Case series are descriptions of a cluster of cases with detailed exposure, clinical and outcome data without controls. Such case series description has led to the identification of AIDS (cluster of Kaposi's sarcoma and of *Pneumocystis carinii* among homosexual men) and Lyme disease (*Borrelia burgdorferi*) (cluster of arthritis in children in Lyme, Connecticut).

Case-control and *cohort studies,* even though important for infectious disease epidemiology, are not discussed here since the methodology is described in detail in earlier chapters (e.g., see chapters ▶Cohort Studies, ▶Case-Control Studies, ▶Modern Epidemiological Study Designs, ▶Epidemiological Field Work in Population-Based Studies, and ▶Exposure Assessment of this handbook).

51.5 Surveillance Issues

Surveillance is the continuous scrutiny of all aspects of occurrence and spread of a disease that are pertinent to effective control (Porta et al. 2008; see also chapter ▶Emergency and Disaster Health Surveillance of this handbook).

The basic activity in surveillance is to identify new cases. Surveillance, both active and passive, is the systematic collection of data pertaining to the occurrence of specific diseases, the analysis and interpretation of these data, and the dissemination of consolidated and processed information to contributors to the program and other interested persons (CDC 2001b). Surveillance has multiple purposes. It provides quantitative data on the magnitude of an illness, documents the distribution throughout the population and the geography leading to information on the natural history of the disease, allows detecting outbreaks, monitors changes in illness patterns, and evaluates the effects of control measures.

51.5.1 Passive Surveillance

In a *passive surveillance system*, the surveillance agency has devised and put a system in place. After the placement, the recipient waits for the provider of care to report. *Passive case detection* has been used for mortality and morbidity data for decades throughout the world. Many countries have an epidemiology section in the health department that is charged with centralizing the data in a national disease surveillance system collecting mortality and morbidity data.

In theory, a passive surveillance system provides a thorough coverage through space and time and gives a thorough representation of the situation. Practically, compliance with reporting is often irregular and incomplete. In fact, the main flaws in passive case detection are incomplete reporting and inconsistencies in case definitions.

The main advantages are the low cost of such a program and the sustained collection of data over decades. The purpose is to produce routine descriptive data on communicable diseases, generate hypotheses, and prompt more elaborate epidemiological studies designed to evaluate prevention activities.

Some conditions must be met to maximize compliance with reporting:

1. Make reporting easy: provide easy to consult lists of reportable diseases, provide prestamped cards for reporting, and provide telephone or fax reporting facilities.
2. Do not require extensive information: name, age, sex, residence, and diagnosis. Some diseases may include data on exposure, symptoms, method of diagnosis, etc.
3. Maintain confidentiality and assure reporters that confidentiality will be respected.
4. Convince reporters that reporting is essential: provide feedback; show how the data are used for better prevention.

Case definitions are important to ensure that data are consistent over time and multiple jurisdictions. On a global basis as well as for pandemics (e.g., the novel H1N1 flu pandemic in 2009), this data consistency is achieved by adhering to WHO case definitions for the respective infectious disease. In the USA, case definitions are regularly updated and published by the Centers for Disease Control and Prevention (CDC) in the Morbidity and Mortality Weekly Report (MMWR) (CDC 1997).

Table 51.8 Hepatitis A case reporting by physicians' specialty and by active/passive sample category, Kentucky, 1983 (Hinds et al. 1985)

	Active sample[a]			Passive sample[b]		
Specialty	*N*	Cases	Rate[c]	*N*	Cases	Rate[c]
General practice/family practice	71	4	5.6	73	2	2.7
Pediatrics	74	7	9.5	71	3	4.2
Internal medicine	71	3	4.2	72	0	0.0
All[d]	216	14	6.5	216	5	2.3

[a] Samples were obtained through weekly phone calls to health-care providers (HCPs)
[b] Samples were sent to the health department without prior phone calls to HCPs
[c] Cases per 100 physicians
[d] Active sample/passive sample rate ratio, adjusted for specialty = 2.8 (95% CI: 1.1–7.2)

Confidentiality of data is essential, particularly for those reporting health-care providers who are subject to very strict confidentiality laws. Any suspicion of failure of maintaining secure data would rapidly ruin a passive surveillance program.

51.5.2 Active Surveillance

In an *active surveillance system*, the recipient will actually take some action to identify the cases. In an active surveillance program, the public health agency organizes a system by searching for cases or maintaining a periodic contact with providers. Regular contacting boosts the compliance of the providers. Providers are health agencies, but also as in passive case detection, there may be day-care centers, schools, long-term care facilities, summer camps, resorts, and even the public involved in reporting diseases to the public health agency.

51.5.2.1 Active Surveillance Through Interaction with Providers

The agency takes the step to contact the health providers (all of them or a carefully selected sample) and requests reports from them at regular intervals. Thus, no reports are missing.

Active surveillance has several advantages:

- It allows the collection of more information. A provider sees that the recipient agency is more committed to surveillance and is therefore more willing to invest more time her/himself.
- It allows direct communication and opportunities to clarify definitions or any other problems that may have arisen.

Active surveillance provides much better and more uniform data than passive case detection (Table 51.8). Active case detection is much more expensive; however, for certain diseases such as Hepatitis A virus (HAV), the benefit normally outweighs the cost. Based on 9 HAV cases and 38 contacts, the total costs for active

Table 51.9 Costs and benefits of a 22-week active surveillance program for Hepatitis A, Kentucky 1983 (Hinds et al. 1985)

Costs		Benefits	
Activity	Dollar estimate	Activity	Dollar estimate
Central office surveillance		Medical costs averted[b]	$5,273
Personnel	$3,764	Indirect costs averted[b]	$8,748
Telephone	$535		
Local health offices[a]			
Contact tracing			
Personnel	$647		
Telephone	$149		
Travel	$31		
Contact prophylaxis			
Personnel	469		
Immune Serum Globulin (ISG)	21		
Total	$5,616		$14,021

[a] Costs of tracing and prophylaxis of 38 additional active surveillance-associated Hepatitis A contacts

[b] Based on 7 Hepatitis A cases prevented among 38 contacts of 9 additional Hepatitis A cases identified by active surveillance. Indirect costs are primarily due to productivity losses

surveillance were estimated to be $5,616; however, the benefits (medical and indirect costs) of 7 HAV cases prevented in among the 38 contacts were $14,021 (Table 51.9).

51.5.2.2 Active Surveillance Through Active Case Detection

Active surveillance systems are usually designed when a passive system is deemed insufficient to accomplish the goals of disease monitoring. This type of surveillance is reserved for special programs, usually when it is important to identify every single case of a disease. Active surveillance is implemented in the final phases of an eradication program. Best examples are the smallpox and poliomyelitis eradication programs and African guinea worm eradication program in some selected countries. Active surveillance is also the best approach in epidemic or outbreak investigations to elicit all cases.

In the smallpox eradication program, survey agents visited providers, asked about suspected cases, and actually investigated each suspected case. In the global polio eradication program which was launched in 1988, all cases of acute flaccid paralysis were investigated.

Thanks to the distribution of water filters, education about the transmission of the parasite, as well as enhanced active surveillance for guinea worm (*Dracunculus medinensis*) there are only five African countries left where dracunculiasis is still endemic. The disease might be eliminated by 2015 which would make guinea worm the first parasite to be eradicated.

51.5.3 Syndromic Surveillance

With increasing concerns about infectious disease outbreaks caused by bioterrorism or emerging infectious agents, it became important to detect health events (illnesses) before final diagnosis or laboratory confirmation. The assumption is that early detection will lead to better prevention. Timeliness and validity of the information are the two most important factors in a successful syndromic surveillance system.

In a syndromic surveillance system, the data collected is not about diagnoses but about indicators of the early stages of an outbreak. Requests for laboratory tests may be part of a syndromic surveillance system, while results of lab tests that may take hours or days would be considered in a passive surveillance system. Other examples of data that may be used are syndromes elaborated from the chief complaints from emergency department records, clinical impressions on ambulance worksheet, prescription filled, retail drug and product purchases, and school or work absenteeism (Buehler et al. 2004).

Framework for evaluating public health surveillance systems for early detection of electronic reporting of data is instrumental in obtaining a rapid transfer of data which is essential for early detection. Statistical tools for pattern recognition and aberration detection are necessary to identify subtle outbreak patterns.

51.5.4 Case Register

A case register is a complete list of all the cases of a particular disease in a definite area over a certain time period. Registers are used to collect data on infections over long periods of time. Registers should be population based, detailed, and complete. A register will show an unduplicated count of cases. They are especially useful for long-term diseases, diseases that may relapse or recur, and diseases for which the same cases will consult several providers and therefore would be reported on more than one occasion.

Case registers contain identifiers, locating information, disease, treatment, outcome, and follow-up information as well as contact management information. They are an excellent source of information for epidemiological studies. In disease control, case registers are indispensable tools for follow-up of chronic infectious diseases such as tuberculosis and leprosy.

The contents and quality of a case register determine its usefulness. It should contain:

- Patient identifiers with names (all names), age, sex, place and date of birth, and complete address with directions on how to reach the patient
- Name and address of a “stable” relative that knows the patient’s whereabouts
- Diagnosis information with disease classification and brief clinical description (short categories are better than detailed descriptions)
- Degree of infectiousness (bacteriological, serological results)
- Circumstances of detection

- Initial treatment and response with specific dose, notes on compliance, side effects, and clinical response
- Follow-up information with clinical response, treatment regimen, compliance, and side effects
- Locating information (for some diseases, contact information is also useful)

Updating a register is a difficult task. It requires cooperation from numerous persons. Care must be taken to maintain the quality of data. It is important to only request pertinent information for program evaluation or information that would remind users to collect data or to perform an exam. For example, if compliance is often a neglected issue, include a question on compliance. Further details concerning the use of registries in general are given in chapter ▶Use of Health Registers of this handbook.

51.5.5 Sentinel Disease Surveillance

For sentinel disease surveillance, only a sample of health providers is used. The sample is selected according to the objectives of the surveillance program. Providers most likely to serve the population affected by the infection are selected; for example, child health clinics and pediatricians should be selected for surveillance of childhood diseases. A sentinel system allows cost reduction and is combined with active surveillance.

A typical surveillance program for influenza infections includes a selected number of general practitioners who are called every week to obtain the number of cases with influenza-like Illness (ILI) presented to them (Fig. 51.7). This

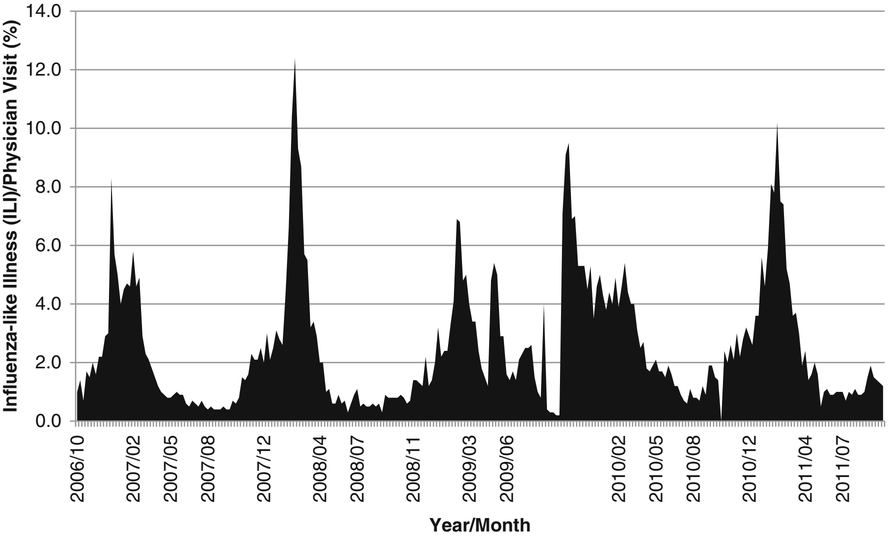


Fig. 51.7 Influenza sentinel surveillance, Louisiana 2006–2011

program may include the collection of samples for viral cultures or other diagnostic techniques. Such a level of surveillance would be impossible to maintain on the national level.

51.5.6 Evaluation of a Surveillance System

Surveillance systems are evaluated on the following considerations (CDC 2001b):

- *Usefulness*: Some surveillance systems are routine programs that collect data and publish results; however, it appears that they have no useful purpose – no conclusions are reached, no recommendations are made. A successful surveillance system would provide information used for preventive purposes.
- *Sensitivity* or the ability to identify every single case of disease is particularly important for outbreak investigations and eradication programs.
- *Predictive value positive* (PVP) is the proportion of reported cases that actually have the health-related event under surveillance. Low PVP values mean that non-cases might be investigated, outbreaks may be exaggerated, or pseudo-outbreaks may even be investigated. Misclassification of cases may corrupt the etiological investigations and lead to erroneous conclusions. Unnecessary interventions and undue concern in the population under surveillance may result.
- *Representativeness* ensures that the occurrence and distribution of cases accurately represent the real situation in the population.
- *Simplicity* is essential to gain acceptance, particularly when relying on outside sources for reporting.
- *Flexibility* is necessary to adapt to changes in epidemiological patterns, laboratory methodology, operating conditions, funding, or reporting sources.
- *Data quality* is evaluated by the data completeness (blank or unknown variable values) and validity of data recorded (see also chapter ▸Quality Control and Good Epidemiological Practice of this handbook).
- *Acceptability* is shown in the participation of providers in the system.
- *Timeliness* is more important in surveillance of epidemics.
- *Stability* refers to the reliability (i.e., the ability to collect, manage, and provide data properly without failure) and availability (the ability to be operational when it is needed) of the public health surveillance system.

51.5.7 Elements of a Surveillance System

The major elements of a surveillance system as summarized by the WHO are mortality registration, morbidity reporting, epidemic reporting, laboratory investigations, individual case investigations, epidemic field investigations, surveys, animal reservoir and vector distribution studies, biologicals and drug utilization, and knowledge of the population and the environment. Traditional surveillance methods rely on counting deaths and cases of diseases. However, these data represent only a small part of the global picture of infectious disease problems.

51.5.7.1 Mortality Registration

Mortality registration was one of the first elements of surveillance implemented. The earliest quantitative data available on infectious disease is about mortality. The evolution of tuberculosis in the USA, for example, can only be traced through its mortality. Mortality data are influenced by the occurrence of disease but also by the availability and efficacy of treatment. Thus, mortality cannot always be used to evaluate the trend of disease occurrence.

51.5.7.2 Morbidity Reporting

Reporting of infectious diseases is one of the most common requirements around the world. A list of notifiable diseases is established on a national or regional level. The numbers of conditions vary; it ranges usually from 40 to 60 conditions. In general, a law requires that health facility staff, particularly physicians and laboratories, report these conditions with guaranteed confidentiality. It is also useful to have other non-health-related entities report suspected communicable diseases such as day-care centers, schools, restaurants, long-term care facilities, summer camps, and resorts. Regulations on mandatory reporting are often difficult to enforce. Voluntary compliance by the institution's personnel is necessary. Reporting may be done in writing, by phone, or electronically in the most advanced system. Since most infectious diseases are confirmed by a laboratory test, reporting by the laboratory may be more reliable. The advantage of laboratory reporting is the ability to computerize the reporting system. Computer programs may be set up to automatically report a defined set of tests and results.

For some infectious diseases, only clinical diagnoses are made. These syndromes may be the consequences of a large number of different microorganisms for which laboratory confirmation is impractical.

When public or physician attention is directed at a specific disease, reporting may be biased. When there is an epidemic or when the press focuses on a particular disease, patients are more prone to look for medical care and physicians are more likely to report. Reporting rates were evaluated in several studies. In the USA, studies show report rates of 10% for viral Hepatitis, 32% for *Hemophilus influenzae*, 50% for meningococcal meningitis, and 62% for shigellosis.

51.5.7.3 Morbidity Case Definition

It is important to have a standardized set of definitions available to providers. Without standardized definitions, a surveillance system may be counting different entities from one provider to another. The variability may be such that the epidemiological information obtained is meaningless.

Most case definitions in infectious disease epidemiology are based on *laboratory tests*; however, some clinical syndromes such as toxic shock syndrome do not have confirmatory laboratory tests. Most case definitions include a brief *clinical description* useful to differentiate active disease from colonization or asymptomatic infection. Some diseases are diagnosed based on epidemiological data. As a result, many case definitions for childhood vaccine preventable diseases and foodborne diseases include epidemiological criteria (e.g., exposure to probable or confirmed cases of disease or to a point source of infection). In some instances, the anatomic site of infection may be important; for example, respiratory diphtheria is notifiable, whereas cutaneous diphtheria is not (CDC 1997).

Cases are classified as a confirmed case, a probable, or a suspected case. An epidemiologically linked case is a case in which (1) the patient has had contact with one or more persons who either have/had the disease or have been exposed to a point source of infection (including confirmed cases) and (2) transmission of the agent by the usual modes is plausible. A case may be considered epidemiologically linked to a laboratory-confirmed case if at least one case in the chain of transmission is laboratory confirmed. Probable cases have specified laboratory results that are consistent with the diagnosis yet do not meet the criteria for laboratory confirmation. Suspected cases are usually cases missing some important information in order to be classified as a probable or confirmed case.

Case definitions are not diagnoses. The usefulness of public health surveillance data depends on its uniformity, simplicity, and timeliness. Case definitions establish uniform criteria for disease reporting and should not be used as the sole criteria for establishing clinical diagnoses, determining the standard of care necessary for a particular patient, setting guidelines for quality assurance, or providing standards for reimbursement. Use of additional clinical, epidemiological, and laboratory data may enable a physician to diagnose a disease even though the formal surveillance case definition may not be met.

51.5.7.4 Data for Which Stage of Disease Should Be Collected? The Morbidity Iceberg

Surveillance programs collect data on the overt cases diagnosed by the health-care system. However, these cases may not be the most important links in the chain of transmission. Cases reported are only the tip of the iceberg (see Fig. 51.8). They may not at all be representative of the true endemicity of an infectious disease.

There is a continuous process leading to an infectious disease: exposed, colonized, incubating, sick, clinical form, convalescing, and cured. Even among those who have overt disease, there are several disease stages that may not be included in a surveillance system:

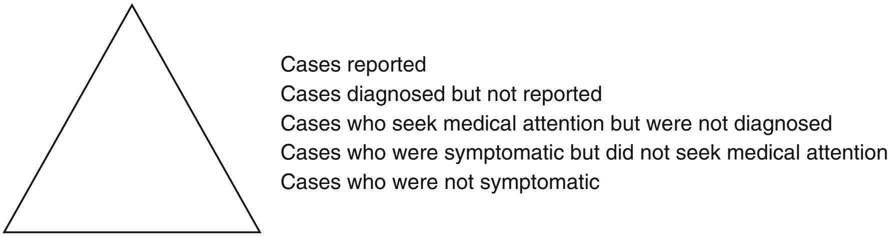

Cases reported
Cases diagnosed but not reported
Cases who seek medical attention but were not diagnosed
Cases who were symptomatic but did not seek medical attention
Cases who were not symptomatic

Fig. 51.8 The iceberg concept

- Some have symptoms but do not seek medical attention.
- Some do get medical attention but do not get diagnosed or get misdiagnosed.
- Some get diagnosed but do not get reported.

Infectious disease cases play different roles in the epidemiology of an infectious disease; some individuals are the indicators (most symptomatic), some are the reservoir of microorganisms (usually asymptomatic, not very sick), some are amplifiers (responsible for most of the transmission), and some are the victims (those who develop severe long-term complications). Depending on the specific disease and the purpose of the surveillance program, different disease stages should be reported. For example:

- In a program to prevent rabies in humans exposed to a suspect rabid animal (usually a bite) needs to be reported. At the stage where the case is a suspect, prevention will no longer be effective.
- For bioterrorism events, reporting of suspects is of paramount importance to minimize consequences. Waiting for confirmation causes too long of a delay. In the time necessary to confirm cases, opportunities to prevent coinfections may be lost, and secondary cases may already be incubating, depending on the transmissibility of the disease.
- Surveillance for West Nile viral infections best rests on the reporting of neuroinvasive disease. Case reports of neuroinvasive diseases are a better indicator than West Nile infection or West Nile fever cases that are often benign, go undiagnosed, and are reported haphazardly.
- For gonorrhea, young males are the indicators because of the intensity of symptoms. Young females are the main reservoir because of the high proportion of asymptomatic infections. Females of reproductive age are the victims because of pelvic invasive disease (PID) and sterility.
- A surveillance program for Hepatitis B that only would include symptomatic cases of Hepatitis B could be misleading. A country with high transmission of Hepatitis B from mother to children would have a large proportion of infected newborns becoming asymptomatic carriers and a major source of infection during their lifetime. Typically in countries with poor reporting of symptomatic Hepatitis, the reporting of acute cases of Hepatitis B would be extremely low in spite of high endemicity which would result in high rates of chronic Hepatitis and hepatic carcinoma.

51.5.7.5 Individual Cases or Aggregate Data?

Most morbidity reporting collects data about individual cases. Reporting of individual cases includes demographic and risk factor data which are analyzed for descriptive epidemiology and for implementation of preventive actions. For example, any investigation leading to contact identification and prophylaxis requires a start from individual cases.

However, identification of individuals may be unnecessary and aggregate data sufficient for some specific epidemiological purposes. Monitoring an influenza epidemic, for example, can be done with aggregate data. Obtaining individual case information would be impractical since it would be too time consuming to collect detailed demographics on such a large number of cases. Aggregate data from sentinel sites consists of a number of influenza-like illnesses by age group and the total number of consultants or the total number of "participants" to be used as denominators. Such data is useful to identify trends and determine the extent of the epidemic and geographical distribution.

Collection of aggregate data of the proportion of school children by age group and sex is a useful predictive tool to identify urinary schistosomiasis endemic areas (Lengeler et al. 2000) without having to collect data on individual school children.

51.5.7.6 Investigations of Cases, Outbreaks, Epidemics, and Surveys

Epidemics of severe diseases are almost always reported. This is not the case for epidemics of milder diseases such as rashes or diarrheal diseases. Many countries do not want to report an outbreak of disease that would cast a negative light on the countries. For example, many countries that are tourism dependent do not report cholera or plague cases. Some countries did not report AIDS cases for a long time.

Case investigations are usually not undertaken for individual cases unless the disease is of major importance such as hemorrhagic fever, polio, rabies, yellow fever, any disease that has been eradicated, and any disease that is usually not endemic in the area.

Outbreaks or changes in the distribution pattern of infectious diseases should be investigated, and these investigations should be compiled in a comprehensive system to detect trends. While the total number of infectious diseases may remain the same, changes may occur in the distribution of cases from sporadic to focal outbreaks. For example, the distribution of WNV cases in Louisiana shifted from mostly focal outbreaks in the first year (2002) the West Nile virus arrived in the state to mostly sporadic cases the following years (2003–2004) (Fig. 51.9).

Surveys are a very commonly used tool in public health, particularly in developing countries where routine surveillance is often inadequate (see chapter ▶Epidemiology in Developing Countries of this handbook). Survey data needs to be part of a comprehensive surveillance database. One will acquire a better picture from one or a series of well-constructed surveys than from poorly collected surveillance data. Surveys are used in control programs designed to control major endemic diseases: spleen and parasite surveys for malaria, parasite in urine and stools for schistosomiasis, clinical surveys for leprosy or guinea worm disease, and skin test surveys for tuberculosis.

51.5.7.7 Surveillance of Microbial Strains

Surveillance of microbial strains is designed to monitor, through active laboratory-based surveillance, the bacterial and viral strains isolated. Examples of these systems are:

- In the USA, the *PulseNet program* is a network of public health laboratories that performs DNA fingerprinting of bacteria causing foodborne illnesses (Swaminathan et al. 2001). Molecular subtyping methods must be standardized to allow comparisons of strains and the building of a meaningful data bank.

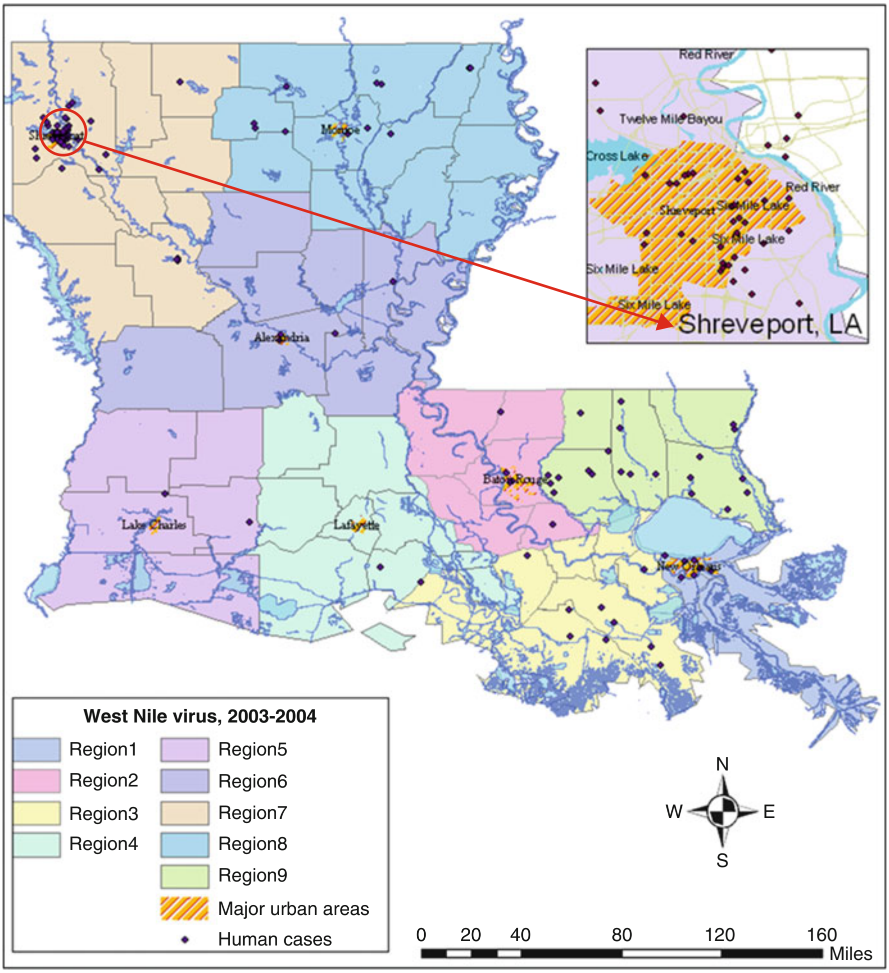


Fig. 51.9 Human West Nile neuroinvasive disease cases in Louisiana, 2003–2004

The method used in PulseNet is pulse field gel electrophoresis (PFGE). The use of standardized subtyping methods has allowed isolates to be compared from different parts of the country, enabling recognition of nationwide outbreaks attributable to a common source of infection, particularly those in which cases are geographically separated.

- The US *National Antimicrobial Resistance Monitoring System* (NARMS) for enteric bacteria is a collaboration between CDC, participating state and local health departments, and the US Food and Drug Administration (FDA) to monitor antimicrobial resistance among foodborne enteric bacteria isolated from humans. NARMS data are also used to provide platforms for additional studies including field investigations and molecular characterization of resistance determinants and to guide efforts to mitigate antimicrobial resistance (CDC 2006).

- Monitoring of antimicrobial resistance is routinely done by requiring laboratories to either submit all or a sample of their bacterial isolates.

51.5.7.8 Surveillance of Animal Diseases

Surveillance for zoonotic diseases should start at the animal level, thus providing early warning for impending increases of diseases in the animal population.

- Rabies surveillance aims at identifying the main species of animals infected in an area, the incidence of disease in the wild animals, and the prevalence of infection in the asymptomatic reservoir (bats). This information will guide preventive decisions made when human exposures do occur.
- Malaria control entomologic activities must be guided by surveillance of *Anopheles* mosquito populations, their biting activities, and *Plasmodium* infection rates in the *Anopheles* mosquitoes.
- Infection rates in wild birds, infection in sentinel chickens, and horse encephalitis are all part of West Nile encephalitis surveillance. These methods provide an early warning system for human infections.
- The worldwide surveillance for influenza is the best example of the usefulness of monitoring animals prior to spread of infection in the human population. Influenza surveillance programs aim to rapidly obtain new circulating strains to make timely recommendations about the composition of the next vaccine. The worldwide surveillance priority is given to the establishment of regular surveillance and investigation of outbreaks of influenza in the most densely populated cities in key locations, particularly in tropical or other regions where urban markets provide opportunities for contacts between humans and live animals (Snacken et al. 1999).

51.5.7.9 Rationale of Selecting Diseases for Surveillance Purposes

The rationale for selecting infectious diseases and an appropriate surveillance method is based on the goal of the preventive program. Table 51.10 shows a few examples of different surveillance methods based on the disease and the objectives of the surveillance.

51.6 Outbreak Investigations

Outbreaks of acute infectious diseases are common, and investigations of these outbreaks are an important task for public health professionals, especially epidemiologists. In 2001, a total of 1,238 foodborne outbreaks with 25,035 cases involved were reported in the USA (CDC 2004) with norovirus being the most common confirmed etiological agent associated with these outbreaks (see Table 51.11).

Table 51.10 Examples of different surveillance methods based on the disease and the objectives of the surveillance system

Disease	Objectives	Surveillance method
Anthrax	Limit bioterrorism event	Active or passive syndromic surveillance
Antibiotic resistance	Description	Active laboratory reporting of antibiograms
Aseptic meningitis	Sentinel event for West Nile Identification of outbreak	Passive surveillance by health-care providers
Gonorrhea	Description of epidemic Treatment of cases	Passive case detection by health-care practitioners Systematic screening of young females (family planning, prenatal, student health services, etc.)
Hepatitis B	Description of endemicity	Survey of representative groups
Hepatitis B	Prevention of perinatal transmission	Screening of pregnant women
Influenza	Quantify endemicity	Sentinel surveillance with aggregate data from physicians’ offices, emergency departments, nursing homes, and schools
Poliomyelitis	Identification of residual cases before complete eradication	Active surveillance of acute flaccid paralysis
Rabies	Prevent human cases	Passive reporting of exposure to potentially rabid animals
Methicillin-resistant *Staphylococcus aureus* (MRSA)	Provide information for management of suspected staphylococcal infections	Active laboratory surveillance of aggregate data on proportion of staphylococci resistant to methicillin
Vancomycin-resistant *Staphylococcus aureus* (VRSA)	Identification of an emerging infection	Laboratory submission of specimens
Tuberculosis	Description of endemicity Case management	Case register
West Nile	Early warning for public and mosquito control	Passive reporting of dead birds by the public, passive reporting of encephalitic horses, sentinel chicken serology, survey of wildlife by serological methods
West Nile	Description of endemicity	Passive and active case finding of neuro-invasive disease

Table 51.11 Confirmed etiological agents of foodborne outbreaks in the USA in 2001

Etiology	Number of outbreaks
Bacillus cereus	5
Brucella spp.	1
Campylobacter spp.	16
Clostridium botulinum	3
Clostridium perfringens	30
Enterohemorrhagic *Escherichia coli*	4
Enterohemorrhagic *Escherichia coli* O157:H7	16
Enterotoxigenic *Escherichia coli*	2
Listeria monocytogenes	1
Salmonella spp.	112
Shigella spp.	15
Staphylococcus aureus	23
Vibrio spp.	4
Yersinia enterocolitica	3
Total bacteria	**235**
Ciguatera	23
Histamine	10
Other chemical	1
Scombroid	18
Total chemical	**52**
Cyclospora cayetanensis	2
Giardia lamblia	1
Trichinella spp.	2
Total parasitic	**5**
Hepatitis A	6
Norovirus	150
Total viral	**156**

Source: CDC Foodborne Outbreak Response and Surveillance Unit (2004)

Outbreaks or epidemics are defined as the number of disease cases above what is normally expected in the area for a given time period. Depending on the disease, it is not always known if the case numbers are really higher than expected and some outbreak investigations can reveal that the reported case numbers did not actually increase. The nature of a disease outbreak depends on a variety of circumstances, most importantly the suspected etiological agent involved, the disease severity or case fatality rate, population groups affected, media pressure, political inference, and investigative progress. There are certain common steps for outbreak investigations as shown in Table 51.12. However, the chronology and priorities assigned to each phase of the investigation have to be decided individually, based on the circumstances of the suspected outbreak and information available at the time.

For example, in 2002, 21 outbreaks of acute gastroenteritis on cruise ships with travel destinations outside the USA were reported to the CDC (CDC 2002). In only

Table 51.12 Fourteen steps in outbreak investigations

#	Step
1	"Outbreak" detected based on initial report or analysis of surveillance data
2	Collect basic numbers and biological specimens
3	Investigate or not?
4	Think prevention first
5	Get information on the disease or condition
6	Sometimes numbers do not count
7	Is the increase real or artificial?
8	Verify the diagnosis
9	Prepare a case definition
10	Put the information in a database
11	Find additional cases
12	Basic descriptive epidemiology (time, place, and person)
13	Hypothesis testing and measures of association
14	Final report and communications

5 of these outbreaks, about 1,400 persons, with an average 280 cases per cruise, had symptoms of viral acute gastroenteritis. Norovirus outbreaks begin usually as a food- or waterborne disease but often continue because of the easy person-to-person transmission in a closed environment and low infectious dose (100 viral particles can be infectious) (CDC 2001a).

51.6.1 Basic Steps in Outbreak Investigations

51.6.1.1 The Initial Report

The original report can originate from very different sources. Examples are:

- A physician is calling the local or state health department about an increase of number of patients seen and diagnosed with a specific disease.
- A high number of patients with similar signs and symptoms are showing up in the emergency room.
- A school principal or day-care owner is reporting a high number of absent students.
- A nursing home health-care professional is seeing a lot of residents with gastrointestinal illnesses.
- A person is complaining to the health department that she/he got sick after eating at a certain restaurant.

Another way to detect an increase of cases is if the surveillance system of reportable infectious diseases reveals an unusually high number of people with the same diagnosis over a certain time period at different health-care facilities.

Outbreaks of benign diseases like self-limited diarrhea are often not detected because people are not seeking medical attention and therefore medical services are not aware of them. Furthermore, early stages of a disease outbreak are often undetected because single cases are diagnosed sporadically. It is not until a certain threshold is passed that it becomes clear that these cases are related to each other through a common exposure or secondary transmission.

Depending on the infectious disease agent, there can be a sharp or a gradual increase of number of cases. It is sometimes difficult to differentiate between sporadic cases and the early phase of an outbreak. In the 2001 St. Louis encephalitis (SLE) outbreak in Louisiana, the number of SLE cases increased from 9 to 18 between weeks 1 and 2, and then the numbers gradually decreased over the next 9 weeks to a total of 63 cases (Jones et al. 2002).

51.6.1.2 Basic Information

After the initial report is received, it is important to collect and document basic information: Contact information of persons affected, a good and thorough event description, names and diagnosis of hospitalized persons (and depending on the presumptive diagnosis their underlying conditions and travel history), laboratory test results, and other useful information to get a complete picture and to confirm the initial story of the suspected outbreak. It also might be necessary to collect more biological specimens such as food items and stool samples for further laboratory testing.

51.6.1.3 Decision to Investigate

On the one hand, based on the collected information, the decision to investigate must be made. It may not be worthwhile to start an investigation if there are only a few people who fully recovered after a couple of episodes of a self-limited, benign diarrhea. Other reasons not to investigate might be that this type of outbreak occurs regularly every summer or that it is only an increase in number of reported cases which are not related to each other.

On the other hand, however, there should be no time delay in starting an investigation if there is an opportunity to prevent more cases or the potential to identify a system failure which can be caused, for example, by poor food preparation in a restaurant or poor infection control practices in a hospital or to prevent future outbreaks by acquiring more knowledge of the epidemiology of the agent involved. Additional reasons to investigate include the interest of the media, politicians, and the public in the disease cluster and the pressure to provide media updates on a regular basis. Another fact to consider is that outbreak investigations are good training opportunities for newly hired epidemiologists.

Sometimes lack of data and lack of sufficient background information make it difficult to decide early on if there is an outbreak or not. The best approach then is to assume that it is an outbreak until proven otherwise.

51.6.1.4 Prevention Comes First

Prevention of more cases is the most important goal in outbreak investigations, and therefore a rapid evaluation of the situation is necessary. If there are precautionary measures to be recommended to minimize the impact of the outbreak and the spread to more persons, they should be implemented before a thorough investigation is completed. Most likely control measures implemented by public health professionals in foodborne outbreaks are:

- Recall or destruction of contaminated food items
- Restriction of infected food handlers from food preparation
- Correction of any deficiency in food preparation or conservation

51.6.1.5 Natural History

After taking immediate control measures, the next step is to know more about the epidemiology of the suspected agent. The most popular books for public health professionals include the "Red Book" (American Academy of Pediatrics 2006), the "Control of Communicable Diseases Manual" from the American Public Health Association (APHA 2008), or other infectious disease epidemiology books as well as the CDC website (www.cdc.gov). If the disease of interest is a reportable disease or a disease where surveillance data are available, baseline incidence rates can be calculated. Then a comparison is made to determine if the reported numbers constitute a real increase or not. Furthermore, the seasonal and geographical distribution of the disease is important as well as the knowledge of risk factors. Many infectious diseases show a seasonal pattern such as rotavirus or *Neisseria meningitides*. For example, in suspected outbreaks where cases are associated with raw oyster consumption, the investigator should know that in the US Gulf states, *Vibrio* cases increase in the summer months because the water conditions are optimal for the growth of the bacteria in water and in seafood. This kind of information will help to determine if the case numbers show a true increase and if it seems likely to be a real outbreak.

51.6.1.6 Number of Cases

For certain diseases, numbers are not important. Depending on the severity of the disease, its transmissibility, and its natural occurrence, certain diseases should raise a red flag for every health professional, and even a single case should warrant a thorough public health investigation. For example, a single confirmed case of a rabid dog in a city (potential dog-to-dog transmission within a highly populated area), a case of dengue hemorrhagic fever, or a presumptive case of smallpox would immediately trigger an outbreak investigation.

51.6.1.7 Artifact

Sometimes an increase of case numbers is artificial and not due to a real outbreak. In order to differentiate between an artificial and a natural increase in numbers, the following changes have to be taken into consideration:

- Alterations in the surveillance system
- A new physician who is interested in the disease and therefore more likely to diagnose or report the disease
- A new health officer strengthening the importance of reporting
- New procedures in reporting (from paper to web-based reporting)
- Enhanced awareness or publicity of a certain disease that might lead to increased laboratory testing
- New diagnostic tests

- A new laboratory
- An increase in susceptible population such as a new summer camp

51.6.1.8 Misclassification

It is important to be sure that reported cases of a disease actually have the correct diagnosis and are not misdiagnosed. Is there assurance that all the cases have the same diagnosis? Is the diagnosis verified and were other differential diagnoses excluded? In order to be correct, epidemiologists have to know the basis for the diagnosis. Are laboratory samples sufficient? If not, what kind of specimens should be collected to ascertain the diagnosis? What are the clinical signs and symptoms of the patient?

In an outbreak of restaurant-associated botulism in Canada, only the 26th case was correctly diagnosed. The slow progression of symptoms and misdiagnosis of the dispersed cases made it very difficult to link these cases and identify the source of the outbreak (CDC 1985, 1987).

51.6.1.9 Case Definition

The purpose of a case definition is to standardize the identification and counting of the number of cases. The case definition is a standard set of criteria and is not a clinical diagnosis. In most outbreaks, the case definition has components of person, place, and time, such as the following: persons with symptoms of X and Y after eating at the restaurant Z between Date1 and Date2. The case definition should be broad enough to get most of the true cases but not too narrow so that true cases will not be misclassified as controls. A good method is to analyze the data, identify the frequency of symptoms, and include symptoms that are more reliable than others. For example, diarrhea and vomiting are more specific than nausea and headache in the case definition of a food-related illness.

51.6.1.10 Database

What kind of information is necessary to be collected? It is sufficient to have a simple database with basic demographic information such as name, age, sex, and information for contacting the patient. More often, date of reporting and date of onset of symptoms are also important. Depending on the outbreak and the potential exposure or transmission of the agent involved, further variables such as school, grade of student, or occupation in adults might be interesting and valuable.

51.6.1.11 Case Finding

During an outbreak investigation, it is important to identify additional cases that may not have been known or were not reported. There are several approaches:

- Interview known cases and ask them if they know of any other friends or family members with the same signs or symptoms.
- Obtain a mailing list of frequent customers in an event where a restaurant is involved.
- Set up an active surveillance with physicians or emergency departments.
- Call laboratories and ask for reports of suspected and confirmed cases.

Another possibility is to review surveillance databases or to establish enhanced surveillance for prospective cases. Occasionally, it might be worthwhile to include the media for finding additional cases through press releases. However, the utility of that technique depends on the outbreak and the etiological agent; the investigator should always do a benefit risk analysis before involving the media.

51.6.1.12 Descriptive Epidemiology

After finding additional cases, entering them in the database, and organizing them, the investigator should try to get a better understanding of the situation by performing some basic descriptive epidemiology techniques such as sorting the data by time, place, and person. For a better visualization of the data, an epidemic or "epi" curve should be graphed. The curve shows the number of cases by date or time of onset of symptoms. This helps to understand the nature and dynamic of the outbreak as well as to get a better understanding of the incubation period if the time of exposure is known. It also helps to determine whether the outbreak had a single exposure and no secondary transmission (single peak) or if there is a continuous source and ongoing transmission. Figures 51.10 and 51.11 show "epi" curves of two different outbreaks: a foodborne outbreak in a school in Louisiana (Fig. 51.10) and the number of WNV human cases, stratified by clinical diagnosis of fever only and meningoencephalitis, in Louisiana in the 2002 outbreak (Fig. 51.11), respectively.

Sometimes it is useful to plot the cases on a map to get a better idea of the nature and the source of an outbreak. Mapping may be useful to track the spread by water (see John Snow's cholera map) or by air or even a person-to-person transmission. If a contaminated food item was the culprit, food distribution routes with new cases identified may be helpful. Maps, however, should be taken with caution and carefully interpreted. For example, WNV cases are normally mapped by residency but do not take into account that people might have been exposed or bitten

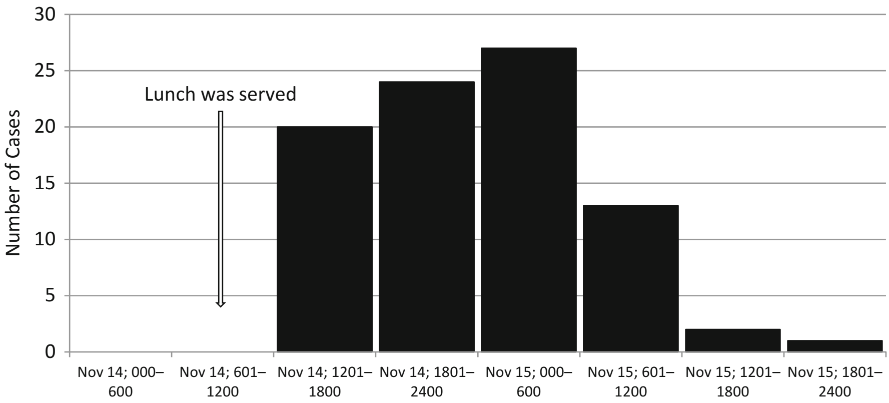


Fig. 51.10 Gastroenteritis outbreak in a school in Louisiana, 2001

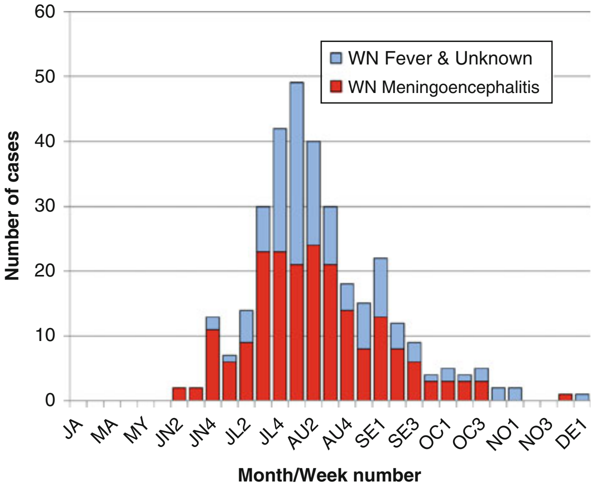


Fig. 51.11 Human West Nile virus cases, Louisiana 2002

by an infective mosquito far away from where they live. For outbreak investigations, spot maps are usually more useful than rate maps or maps of aggregate data.

Depending on the outbreak, it might be useful to characterize the outbreak by persons' demographics such as age, sex, address, occupation, and health status. Are the cases at increased susceptibility or at high risk of infection? These kinds of variables might give the investigator a good idea if the exposure is not yet known. For typical foodborne outbreaks, however, demographic information is not very useful because the attack rates will be independent of age and sex. More details on methods used in descriptive epidemiology are given in chapter ▶Descriptive Studies of this handbook.

51.6.1.13 Hypothesis Forming

Based on the results of basic descriptive epidemiology and the preliminary investigation, some hypotheses should be formulated in order to identify the cause of the outbreak. A hypothesis will be most likely formulated such as "those who attended the luncheon and ate the chicken salad are at greater risk than those who attended and did not eat the chicken salad." It is always easier to find something after knowing what to look for, and therefore a hypothesis should be used as a tool. However, the epidemiologist should be flexible enough to change the hypothesis if the data do not support it. If data clues are leading in another direction, the hypothesis should be reformulated such as "those who attended the luncheon and ate the baked chicken are at greater risk than those who attended and did not eat the baked chicken."

To verify or deny hypotheses, measures of risk association such as the relative risk (*RR*) or the odds ratio (*OR*) have to be calculated (as described in chapters ▶Rates, Risks, Measures of Association and Impact, ▶Cohort Studies, and

►Case-Control Studies of this handbook). The CDC has developed the software program “Epi Info” which is easy to use in outbreak investigations and, even more importantly, free of charge. It can be downloaded from the CDC website (http://www.cdc.gov/epiinfo/). Measures of association, however, should be carefully interpreted; even a highly significant measure of association cannot give enough evidence of the real culprit or the contaminated food item. The measure of association is only as good and valid as the data. Most people have recall problems when asked what they ate, when they ate, and when their symptoms started. Even more biases or misclassifications of cases and controls can hide an association. A more confident answer comes usually from the laboratory samples from both human samples and food items served at time of exposure. Agents isolated from both food and human samples that are identified as the same subtype, in addition to data results supporting the laboratory findings, are the best evidence beyond reasonable doubt.

51.6.1.14 Final Report

As the last step in an outbreak investigation, the epidemiologist writes a final report on the outbreak and communicates the results and recommendations to the public health agency and facilities involved (see Table 51.12). In the USA, public health departments also report foodborne outbreaks electronically to CDC via a secure web-based reporting system, the National Outbreak Reporting System (NORS).

51.6.2 Types of Outbreaks

51.6.2.1 The “Traditional” Foodborne Outbreak

The “traditional” foodborne outbreak is usually a small local event such as family picnic, wedding reception, or other social event and occurs often in a local restaurant or school cafeteria. This type of outbreak is highly local with a high attack rate in the group exposed to the source. Because it is immediately apparent to those in the local group such as the group of friends who ate at the restaurant or the students’ parents, public health authorities are normally notified early in the outbreak, while most of the cases are still symptomatic. Epidemiologists can start early on with their investigation and therefore have a much better chance to collect food eaten and stool samples of cases with gastroenteritis for testing and also to detect the etiological agent in both of them.

In a 2001 school outbreak in Louisiana, 87 persons (67 students and 20 faculty members) (see Fig. 51.10) experienced abdominal cramps after eating at the school’s annual “Turkey Day” the day before. Stool specimens and the turkey with the gravy were both positive for *Clostridium perfringens* with the same pulse field gel electrophoresis (PFGE) pattern (Merlos 2002). The inspection of the school cafeteria revealed several food handling violations such as storing, cooling, and reheating of the food items served. Other than illnesses among food handlers, these types of improper food handling or storage are the most common causes of foodborne outbreaks.

51.6.2.2 New Types of Outbreaks

A different type of outbreak is emerging as the world is getting smaller. In other words, persons and food can travel more easily and faster from continent to continent and so do infectious diseases with them. Foodborne outbreaks related to imported contaminated food items are normally widespread, involving many states and countries, and therefore are frequently identified. In 1996, a large outbreak of *Cyclospora cayetanensis* occurred in 10 US states and Ontario, Canada, and was linked to contaminated raspberries imported from South America. Several hundred laboratory-confirmed cases were reported, most of them immunocompetent persons (CDC 1996).

A very useful molecular tool to identify same isolates from different geographical areas is subtyping enteric bacteria with PFGE. In the USA, the PulseNet database allows state health departments to compare their isolates with other states and therefore increase the recognition of nationwide outbreaks linked to the same food item (Swaminathan et al. 2001).

In a different scenario, a widely distributed food item with low-level contamination might result in an increase of cases within a large geographical area and therefore might be not get detected on a local level. This kind of outbreak might only be detected by chance if the number of cases increased in one location and the local health department alerts other states to be on the lookout for a certain isolate.

Another type of outbreak is the introduction of a new pathogen into a new geographical area as it happened in 1991 when *Vibrio cholerae* was inadvertently introduced in the waters off the Gulf Coast of the United States.

Food can not only be contaminated by the end of the food handling process, that is, by infected food handlers, but also can be contaminated by any event earlier in the chain of food production. In 1996, an outbreak of *Salmonella enteritidis* in a national brand of ice cream resulted in 250,000 illnesses. The outbreak was detected by routine surveillance because of a dramatic increase of *Salmonella enteritidis* in South Minnesota. The cause of the outbreak was a basic failure on an industrial scale to separate raw products from cooked products. The ice-cream premix was pasteurized and then transported to the ice-cream factory in tanker trucks which had been used to haul raw eggs. This resulted in the contamination of the ice cream and subsequent salmonella cases (Hennessy et al. 1996).

51.7 Surveys

Surveys are useful to provide information for which there is no data source or no reliable data source. Surveys are time consuming and are often seen as a last choice to obtain information. However, too often unreliable information is used because it is easily available. For example, any assessment of the *Legionella* problem using passive case detection will be unreliable due to underdiagnosis and underreporting. Most cases of legionellosis are treated empirically as community-acquired pneumonias and are never formally diagnosed.

In developing countries, surveys are often necessary to evaluate health problems since data collected routinely (disease surveillance, hospital records, case registers) are often incomplete and of poor quality. In industrialized nations, although many sources of data are available, there are some circumstances where surveys may be necessary.

Prior to carrying out surveys involving human subjects, special procedures need to be followed. In industrialized countries, a human subject investigation review board has to evaluate the project's value and ethics. In developing countries, however, such boards may not be formalized, but it is important to obtain permission from medical, national, and local political authorities before proceeding.

51.7.1 Survey Methods

Surveys of human subjects are carried out by mail, telephone, personal interviews, and behavioral observations. In infectious diseases, the collection of biological specimens in humans (i.e., blood for serological surveys) or the collection of environmental samples (food, water, environmental surfaces) is very common. Personal interviews and specimen collection require face-to-face interaction with the individual surveyed. These are carried out in offices or by house-to-house surveys.

Non-respondents are an important problem for infectious disease surveys. Those with an infection may be absent from school, may not answer the door, or may be unwilling to donate blood for a serological survey, thus introducing a systematic bias into the survey results.

Since surveys are expensive, they cannot be easily repeated. All field procedures, questionnaires, biological sample collection methods, and laboratory tests should be tested prior to launching the survey itself. Feasibility, acceptability, and reliability can be tested in a small-scale pilot study. More details on survey methods are to be found in chapter ▶Epidemiological Field Work in Population-Based Studies of this handbook.

51.7.2 Sampling

Since surveys are labor intensive, they are rarely carried out on an entire population but rather on a sample. To do a correct sampling, it is necessary to have a sampling base (data elements for the entire population) from which to draw the sample. Examples of sampling bases are population census, telephone directory (for the phone subscriber population), school roster, or a school list. In developing countries, such lists are not often available and may have to be prepared before sampling can start. More information on sampling designs can be found in chapter ▶Epidemiology in Developing Countries of this handbook.

51.7.3 Community Surveys (House-to-House Surveys)

Most community surveys are carried out in developing countries because reliable data sources are rare. The sampling base often ends up to the physical layout of the population. A trip and geographical reconnaissance of the area are necessary. The most common types of surveys undertaken in developing countries are done at the village level; they are based on maps and a census of the village.

In small communities, it is important to obtain the participation of the population. Villagers are often wary of government officials counting people and going from door to door. To avoid misinterpretations and rumors, influential people in the community should be told about the survey. Their agreement is indispensable, and their help is needed to explain the objectives of the survey and particularly its potential benefits. Increasing the knowledge about disease, disease prevention and advancing science are abstract notions that are usually poorly understood or valued by villagers who are, in general, very practical people. If a more immediate benefit can be built into the survey, there will be an increase in cooperation of the population. Incentives such as offering to diagnose and treat an infection or drugs for the treatment of common ailments such as headaches or malaria enhance the acceptance of the survey.

In practically all societies, the household is a primary economic and social unit. It can be defined as the smallest social unit of people who have the same residency and maintain a collective organization. The usual method for collecting data is to visit each household and collect samples or administer a questionnaire.

Medical staff may feel left out or even threatened whenever a medical intervention (such as a survey) is done in their area. A common concern is that people will go to their medical care provider and ask questions about the survey or about specimen collection and results. It is therefore important to involve and inform local medical providers as much as practical.

A rare example of a house-to-house survey in an industrialized nation was carried out in Slidell, Louisiana, for the primary purpose of determining the prevalence of West Nile infection in a southern US focus. Since the goal was to obtain a random sample of serum from humans living in the focus, the only method was a survey of this type. A cluster-sampling design was used to obtain a representative number of households. The area was not stratified because of its homogeneity. Census blocks were grouped so that each cluster contained a minimum of 50 households. The probability of including an individual cluster was determined by the proportion of houses selected in that cluster and the number of persons participating given the number of adults in the household. A quota sampling technique was used, with a goal of enlisting ten participating households in each cluster.

Inclusion criteria included age (at least 12 years of age) and length of residence (at least 2 years). The household would be included only if an adult household resident was present. A standardized questionnaire was used to interview each participant. Information was collected on demographics, any recent febrile illness, knowledge, attitudes, and behaviors to prevent WNV infection and potential exposures to mosquitoes. A serum sample for WNV antibody testing was drawn.

In addition, a second questionnaire regarding selected household characteristics and peridomestic mosquito reduction measures was completed. Informed consent was obtained from each participant, and all participants were advised that they could receive notification of their blood test results if they wished. Institutional review board approvals were obtained.

Logistics for specimen collection, preservation, and transportation to the laboratory were arranged. Interpretation of serological tests and necessary follow-up were determined prior to the survey and incorporated in the methods submitted to the ethics committee.

Sampling weights, consisting of components for block selection, household-within-block selection, and individual-within-household participation, were used to estimate population parameters and 95% confidence intervals (CI). Statistical tests were performed incorporating these weights and the stratified cluster sampling design.

In this survey, 578 households were surveyed (a 54% response rate), including 1,226 participants. There were 23 Immunoglobulin M (IgM) seropositive persons, for a weighted seroprevalence of 1.8% (with a 95% confidence interval of 0.9–2.7%) (Michaels et al. 2005).

51.8 Microbiological and Serological Issues

51.8.1 Case Confirmation

Definitive confirmation of a case relies on the identification of the infectious agent in the patient in some specific body sites. The site is important for agents that may be pathogens or colonizers. For example, identification of *Neisseria meningitidis* in an upper respiratory fluid could be due to colonization, while isolation from cerebrospinal fluid or blood would mean definite invasive disease.

Identification of infectious agents is more frequently done by genotyping of the agent nowadays. The reliability of these identifications depends on the methods being used. Many of the tests are developed "in house." The tests that identify a single gene are less reliable than those that identify several specific genes.

51.8.2 Serological Issue and Seroepidemiology

Serological tests have long been used to diagnose the cause of an infectious disease. In most instances, these methods are reliable, but this is not always so. The old serological methods (agglutination, hemagglutination, complement fixation, etc.) often resulted in a false-positive/false-negative result. Dilutions were used to quantify the reactions. In general, positive reactions at low titers were meaningless, while positivity at high dilutions was indicative of a recent infection. A fourfold increase of positivity (e.g., from a 1:16 to 1:64 dilution) over a 2-week period was usually considered confirmatory based on the dilution factor that meets the criteria for positivity in the laboratory test. A good example for this is *Brucella* sp. (a zoonotic disease where humans are accidental hosts) where a fourfold or greater rise in the *Brucella* agglutination titer between acute (specimen taken while patient is symptomatic) and convalescent (specimen taken while patient is recovering from the disease) serum specimens obtained 2 or more weeks apart is considered positive. Serological diagnosis is often discouraged because of difficulties of collecting follow-up serum samples if the patient has recovered and is not in medical care anymore.

Newer techniques such as enzyme immunoassays (EIA) do quantify the amount of antibodies present, but do not allow for the "fourfold increase."

The serological response to an infection consists of several types of antibodies: IgM at first, Immunoglobulin G (IgG) antibodies later, and also Immunoglobulin A (IgA) antibodies. The usual assumption is that IgM antibodies are produced early (before IgG antibodies) and for a limited length of time (usually 2–3 months). However, the postulate "IgM means recent infection" is not always true. Onset and length of production of IgM antibodies depend on the infectious agent. For West Nile infections, IgM production starts a few days after infection but may last for several years; in fact, 1 year after infection, 40% of the patients are still reported IgM positive by most criteria established by laboratory definitions.

51.8.3 Sensitivity, Specificity, and Predictive Value of a Positive Test

Issues of sensitivity, specificity, and predictive value are particularly relevant to serological testing. The methods are similar to those described in chapter ▶Clinical Epidemiology and Evidence-Based Health Care of this handbook. The predictive value of a positive test (*PVPT*) depends on its sensitivity, specificity, and prevalence. Its formula is the following:

$$PVPT = \frac{PR^{*}SE}{(PR^{*}SE) + [(1 - PR)^{*}(1 - SP)]},$$

where *PR* denotes the prevalence, *SE* the sensitivity, and *SP* the specificity.

It is heavily influenced by the prevalence. Even if the tests have the highest sensitivity and specificity (99% for both in Table 51.13), the predictive value is poor

Table 51.13 Predictive value of a positive test

Sensitivity (%)	Specificity (%)	Prevalence (%)	*PVPT* (%)
99	99	5	83.9
99	99	1	50
99	99	0.1	9
95	95	5	50
95	95	1	16.1
95	95	0.1	1.9

when the prevalence is low. This has implications for case definitions in situations of very low prevalence or disappearing infectious diseases (example of measles and rubella in the USA).

51.9 Nosocomial Infection Epidemiology

Epidemiology plays a major part in prevention programs against nosocomial (hospital-acquired) infections. Surveillance should provide systematic and continuous observations on the occurrence and distribution of nosocomial infections within the hospital population. Surveillance is the focal point for infection control activities. The term surveillance implies that the observational data are regularly analyzed.

Surveillance activities may provide valuable epidemiological data such as the identification of outbreaks, priorities for infection control activities, and the elucidation of important secular trends, such as shifts in microbial pathogens, infection rates, or outcomes of hospital-acquired infection. Surveillance activities provide the additional benefits of increasing the visibility of the infection control team in the hospital during the infection control practitioners' ward rounds and of allowing an opportunity for informal consultation and education for both nurses and physicians.

Ideally, the surveillance of hospital-acquired infection should be a continuous process that consists of the following elements:

1. Definition of categories of infection
2. Systematic case finding and data collection
3. Tabulation of data
4. Analysis and interpretation of data
5. Reporting of relevant infection surveillance data to individuals and groups for appropriate action

51.9.1 Definitions

The use of consistent definitions of nosocomial infection is critical in developing data on endemic infection rates. Definitions must be simple, requiring only clinical information or readily available laboratory data.

51.9.1.1 General Definitions

A nosocomial infection is either:

1. An infection which is acquired during hospitalization and which was not present or incubating at the time of admission or
2. An infection which is acquired in the hospital and becomes evident after discharge from the hospital or
3. A newborn infection which is the result of passage through the birth canal.

An infection is defined as hospital acquired if the patient (1) has an infection, not a simple colonization, (2) was not infected at the time of admission, and (3) had sufficient time to develop infection.

True Infection and Not Colonization or Contamination Infections are accompanied by signs and symptoms of infection (fever, malaise) and in localized infections: swelling due to inflammation, heat, pain, and erythema (tumor, dolor, rubor, or calor). Immunocompromised patients do not show signs of infection as easily as normal patients. Neutropenic patients ($\leq$500 neutrophils/cubic millimeter) show no pyuria, no purulent sputum, little infiltrate, and no large consolidation on chest X-ray. An antibiotic treatment by a physician is a presumption of infection.

No Infection at Time of Admission Several criteria may be used to establish prior negativity: history, symptoms and signs documented at the time of admission, lab tests, and chest X-rays done in the early days in the hospital. Normal physical examination, absence of signs and symptoms, normal chest X-ray, negative culture, and lack of culture are useful.

Sufficient Time to Develop Infection For diseases which have a specific incubation period, the hospital-acquired infection can only develop if the patient has stayed in the hospital for a stay $\geq$ incubation period. Numerous infections do not have well set incubation periods (e.g., staphylococci and *E. coli* infections). However, these infections rarely develop in less than 2 days.

To establish a nosocomial infection meeting the definition criteria, it is sufficient that there is no need to have proof beyond the shadow of a doubt.

51.9.1.2 Specific Definitions

To carry out surveillance, very specific definitions are necessary, not only regarding the major nosocomial infections (surgical site infection, bloodstream infections, pneumonia, and urinary tract infections) but regarding all possible sites of nosocomial infections.

51.9.2 Scope/Strategy of Surveillance

Active surveillance is much more effective than *passive surveillance*. Using active surveillance increases the sensitivity of identifying infections.

51.9.2.1 Case Finding

Case finding can be *retrospective*, *prospective*, or both. Prospective or concurrent surveillance means monitoring the patient during hospitalization. Prospective surveillance may include the post-discharge period. In contrast, retrospective surveillance involves review of the medical record after the patient has been discharged. Prospective surveillance provides increased visibility for infection control personnel and timely analysis of data and feedback to clinical services, but this type of surveillance is more expensive. Retrospective methodology is cheaper to implement but requires more controls to verify how effective the infection control personnel are as follows:

- Patient-based case finding relies on evaluating medical records and doing rounds in hospital wards. It allows assessing risk factors, procedures, and practices related to patient care.
- Laboratory-based surveillance relies on identifying positive cultures for pathogens. Then, further investigations are necessary to verify if this is a health-care facility-associated infection, a community-associated infection, a colonization, or a contamination.

A major issue is determining the scope of surveillance. Choices can include three major strategies: hospital-wide surveillance, surveillance by objective, and limited or targeted surveillance.

51.9.2.2 Hospital-Wide Surveillance

Comprehensive or hospital-wide surveillance implies a continuous surveillance of all patients for all types of nosocomial infections in all hospital wards.

This strategy is time consuming. Efficiency is increased by using "clues" to identify patients whose charts should be reviewed. A hospital-wide surveillance provides a global view of the hospital, but the cost and the labor involved may be prohibitive. Critics of whole-house surveillance argue that collecting and analyzing data may be overwhelming; time may not permit developing objectives for surveillance, and many of the identified infections may not be preventable. A modification of this strategy includes doing hospital-wide surveillance for 1 year, or part of a year.

51.9.2.3 Surveillance by Objectives

Surveillance by objectives focuses on specific outcome objectives defined for surveillance purposes. Levels of surveillance effort are prioritized. Prioritization focuses on types of infections to be prevented, and levels of effort may be adjusted to the relative seriousness of the problem. Considerations in setting these priorities would include morbidity and mortality data, costs of treating infections, length of stay, frequency of occurrence of infection, and percentage of infections that are thought to be preventable. If baseline rates (before objectives are met) are not established, the identification of clusters and epidemics would be difficult. Areas that were not included in the objectives could not be evaluated.

51.9.2.4 Targeted Surveillance

Targeted surveillance can be site specific, unit specific, rotating, or limited to outbreak surveillance.

Site-specific surveillance focuses on specific infection sites such as surgical wounds or urinary tract infections. In contrast to surveillance by objective, this strategy lacks a defined objective. It is flexible, because this strategy can be used concurrently with alternating components such as continuous and rotating surveillance as well as special projects.

Unit-directed surveillance targets specific units or areas with highest risk. Surveillance activities are limited to the areas of highest risks such as intensive care units, burn units, and hematology and oncology units. Targeting critical care and oncology units, for example, would capture the majority of all bloodstream infections. The rate of infection in these units is high, yet a relatively small number of patients are actually treated. This approach may prevent infections in patients at greatest risk.

Rotating surveillance is periodic and systematic surveillance in a given unit for a specific time period. This technique is less time consuming and more cost effective than other forms of surveillance because all areas of the hospital are covered at sequential periodic intervals using careful continuous surveillance. Ideally, rotating surveillance involves an annual, detailed, and directed infection-control evaluation for each hospital unit. One type of rotating surveillance, the prevalence survey, can identify infection control risks; however, it can also miss clusters in areas that are not currently under surveillance.

Outbreak surveillance requires an alert hospital staff who report any unusual cluster of events that, when based on surveillance data, extend beyond threshold units.

Whichever surveillance strategies are selected, they should allow personnel to recognize and workup clusters of infections or events.

51.9.3 Calculating Rates

51.9.3.1 Numerator

The numerators may be the number of infections or the number of patients infected. Decisions must be made on how to count infections caused by multiple organisms at the same site (usually counted as one infection), infections in a patient with a second nosocomial infection, a patient with an extension of another infection, and so forth.

51.9.3.2 Denominators

If incidence rates are warranted, a common denominator is number of patients admitted or discharged. If incidence density rates are calculated, number of hospital days or numbers of device days are usually used. The choice of denominator depends on the purpose of calculating these rates.

Hospital-Wide Nosocomial Infection Rate per 100 Admissions A hospital-wide nosocomial infection rate/100 admissions for a given period (month, quarter, or year) is commonly calculated but has little significance because it does not take into account (1) the risk posed to the patient by procedures (intravenous (IV) lines, urinary catheters, or ventilators) and (2) the severity of the patients' conditions. A small hospital with little use of invasive procedures and relatively healthier patients will have lower rates of infection.

In this rate, a patient who has two infections is actually counted twice. This rate can be calculated as

$$\frac{\text{number of nosocomial infections}^{*}100}{\text{number of patients admitted}}.$$

Hospital-Wide Patient Infection Rate per 100 Admissions Hospital-wide patient infection rate/100 admissions are used to avoid the pitfall of multiple infections in the same patient. This rate may be calculated for a given period: month, quarter, and year. In this rate, a patient with two infections is counted only once:

$$\frac{\text{number of patients infected}^{*}100}{\text{number of patients admitted}}.$$

Patient Infection Rate per 1,000 Hospital Days The risks of infections are much higher in some units of the hospital such as intensive care units (ICUs) and coronary care units. Calculating ward-specific rates is useful to look at the trends in specific units or compare between units. The number of patients admitted to an ICU may be difficult to determine because some patients are admitted in the ward 1 day, spend a few days, be discharged to a regular ward, and after a few days be readmitted into ICU. To avoid the problem posed by the same patient admitted and discharged several times from the ICU to the wards, the rate of infection is expressed in number of patients infected/1,000 hospital days. This rate also takes into account the duration of hospitalization which is a risk factor for nosocomial infection. It allows comparison between wards where duration is different and can be calculated as

$$\frac{\text{number of infections}^{*}100}{\text{number of hospital days}}.$$

Device-Specific Rates and Procedure-Specific Rates The risk of infection is related to the extrinsic risk factors (use of devices such as ventilator, central line intravascular catheter, urinary catheter, surgical operation). To compare the risk associated with these devices or procedures, the following rates are best suited:

$$\text{surgical site infection rate} = \frac{\text{number of surgical site infections}^{*}100}{\text{number of patients operated on}},$$

$$\begin{aligned}&\text{ventilator associated pneumonia rate}\\&= \frac{\text{number of ventilator associated pneumonia}^{*}1{,}000}{\text{number of patients on ventilator days}},\end{aligned}$$

$$\text{catheter related bloodstream infection (BSI) rate} \\ = \frac{\text{number of catheter related BSI}^* 100}{\text{number of patients on IV line days}},$$

where it may be difficult to obtain the number of intravascular line days. Ideally, the number should be line specific, for example, central line (which is a catheter (tube) that is passed through a vein to end up in the thoracic (chest) portion of the vena cava or in the right atrium of the heart) days and peripheral line (which is a catheter (tube) placed into a peripheral vein) days

$$\text{utilization rate} = \frac{\text{number of device days }^* 100}{\text{number of patients days}},$$

where the device utilization rate (DUR) is the proportion of patient days for which a certain device is used. The DUR is specific to a certain device: catheter, IV line, and ventilator. The DUR reflects the amount of devices used and is a reflection of the patient severity.

51.10 Epidemiological Aspects of Infectious Disease Prevention

51.10.1 Antibiotic Resistance

There are an almost daily increasing number of publications on antibiotic resistance creating the impression that the resistance is growing worldwide. However, there is no comprehensive surveillance system for antibiotic sensitivity and no comprehensive database documenting the spread of resistance in the USA or worldwide. One of the best data sources for the USA comes from the *National Nosocomial Infection Surveillance* (NNIS), now the *National Healthcare Safety Network* surveillance which documents sensitivity of nosocomial infections, an estimated 4% of bacterial infections occurring in the USA.

On a very global aspect, antibiotics are still very effective. For many hospitals, antibiotic sensitivity patterns are not very different nowadays than what they were 10 years ago, except for very few pathogens. Reports obtained from the medical literature are not representative of the whole hospitals, and even of the whole “world” of bacteria. Many of the resistance reports from the literature come from single institutions where antibiotic resistance was the consequence of overuse of an antibiotic. Very few reports attempt to compare several institutions. There is no randomized, non-selective, multicentered data to evaluate the scope of resistance and its evolution. Most of the US data is reported from large metropolitan hospitals affiliated with medical schools in the northern USA. These hospitals treat the more severely ill patients who often have been treated unsuccessfully at community hospitals and have probably become resistant during previous attempts at treatment.

The impression is that the most severe cases of resistance are generated in the tertiary care hospitals (these are specialty hospitals dedicated to specific subspecialty care including ICUs). It may well be that resistance is generated in primary (health care was provided by a general practitioner or other health professional) and secondary care facilities (health care provided by hospital clinicians) and those cases who did not respond to antibiotics were referred to tertiary care hospitals.

51.10.1.1 Active Surveillance

The goal of an antibiotic sensitivity active surveillance system is to estimate the proportion of selected bacteria that are resistant to antibiotics by the reporting of laboratory aggregate data. This surveillance system can only monitor a few pathogens. In the USA, the most common pathogens monitored in such programs are methicillin-resistant *Staphylococcus aureus* (MRSA), drug-resistant *Streptococcus pneumoniae* (DRSP), and vancomycin-resistant *Enterococcus* (VRE). Laboratories are asked to report:

1. The total number of drug-resistant or drug intermediate-resistant isolates excluding duplicates (one isolate per patient per month if possible) (numerator)
2. The total number of isolates of the bacterial species of concern (denominator) from a given laboratory for each month

51.10.1.2 Antibiograms

Another approach to establish an antibiotic sensitivity surveillance system is to use the hospital antibiograms. In 2001, the *National Committee on Clinical Laboratory Standards* (NCCLS) now known as the *Clinical and Laboratory Standards Institute* (CLSI) issued guidelines on how to analyze and present cumulative antimicrobial sensitivity test data from antibiotic sensitivity testing performed on health-care facility patients. The data show the percent sensitivity for the first isolate from a patient within an analysis period (generally 1 year), the specimen source, and the total number of isolates tested (minimum ten for each organism to avoid describing sensitivity on a sample of less than ten patients).

The compilation of individual hospital antibiograms over time is useful in monitoring antibiotic sensitivity. The CDC conducted a study to compare data from the resource-intensive active surveillance collection of antibiotic resistance patterns to the data collected using hospital antibiograms. The study found the proportions of drug-resistant isolates from antibiograms were within ten percentage points of those from isolates obtained through active surveillance, thereby providing a relatively simple and accurate way to monitor antibiotic resistance (Van Beneden et al. 2003).

Limitations of hospital antibiograms are that they do not sort out community-acquired infections from nosocomial infections and some laboratories may not thoroughly unduplicate their data, thus giving a picture of a larger number of resistant isolates than it is the case.

51.10.2 Immunization

Epidemiology plays a major role at several stages in immunization programs.

51.10.2.1 At the Development Stage

Once a vaccine has been developed, it has to go through a rigorous process to be recognized as safe and efficacious. Once information on the vaccine composition, manufacturing, stability and sterility, and animal testing results have been submitted for review, the vaccine has to go through preclinical and clinical trials. In the preclinical studies, assays are carried out in animals to determine the humoral and cellular responses, the optimal administration route, the dose-response relationship, and the dosing schedule and the adverse or toxic effects. This is followed by the clinical studies. Phase I studies are intended to determine the efficacious dose and safety of the vaccine in a small number of healthy adults. Phase II studies are more extensive "open-label" prospective cohort studies or small randomized controlled trials on all relevant age groups. Their goal is to establish safety and immunogenicity. Phase III studies are randomized double-blind placebo-controlled vaccine efficacy trials. A comparison is made for the incidence rate of a disease in the standard versus a placebo group. The goal is to confirm the efficacy and obtain a comprehensive list of side effects.

51.10.2.2 At the Implementation Stage

Once in public use, the populations receiving the vaccine are much less controlled than during the trials. The designs of epidemiological studies must be adapted to these new conditions. Descriptive studies, surveys, case-control, and cohort studies are then performed with a goal to evaluate efficacy, side effects, and success of a vaccination campaign. The study of outbreaks among unvaccinated populations becomes a very useful tool to evaluate efficacy.

51.10.2.3 When Vaccine Led to Disappearing Illness: Eradication

Once a vaccine has been widely distributed among the population and the herd immunity is very high, the incidence of disease will decrease until elimination. Epidemiological studies are useful to determine if the widespread use of vaccine has led to suppression of disease with continuation of circulation of the agent or to the total disappearance of the infectious agent. With poliomyelitis, the killed vaccine led to elimination of the disease, but the virus was still circulating. The live oral vaccine on the other hand led to a complete elimination of the circulating virus. Epidemiological methods are instrumental in gathering this evidence.

During the final stages of a disappearing illness, active surveillance and detailed case investigation are necessary to detect every suspect and confirm the diagnosis with a definitive laboratory test (under these circumstances, identification of the infectious agent is preferred to a serological/immunological test). For poliomyelitis eradication, surveillance for acute flaccid paralysis is implemented before declaring a country free of the disease.

51.11 Program Evaluation

Program evaluation is a systematic way to determine if prevention or intervention programs for the infectious disease of interest are effective and to see how they can be improved. It is beyond the scope of this chapter to explain program evaluation in detail; however, there is abundant information available, that is, the CDC's Framework for Program Evaluation in Public Health (CDC 1999) as well as many valuable text books on program evaluation.

Most importantly, evaluators have to understand the program such as the epidemiology of the disease of interest, the program's target population and their risk factors, program activities, and resources. They have to identify the main objectives of the control actions and determine the most important steps. Indicators define the program attributes and translate general concepts into measurable variables. Data are then collected and analyzed so that conclusions and recommendations for the program are evidence based.

Evaluating an infectious disease control program requires a clear understanding of the microorganism, its mode of transmission, the susceptible population, and the risk factors. The following example of evaluation of tuberculosis control shows the need to clearly understand the priorities.

Most of tuberculosis transmission comes from active pulmonary tuberculosis cases that have positive sputum smears (confirmed as *Mycobacterium tuberculosis* on culture). To a lesser extent, smear-negative, culture-positive pulmonary cases are also transmitting the infection. Therefore, priority must be given to find sputum-positive pulmonary cases. The incidence of smear-positive tuberculosis cases is the most important incidence indicator. Incidences of active pulmonary cases and of all active cases (pulmonary and extrapulmonary) are also calculated but are of lesser interest. The proportion of all cases of tuberculosis that are pulmonary versus extrapulmonary, smear-positive culture-positive pulmonary versus culture-positive only pulmonary, or culture-negative pulmonary is used to detect anomalies in case finding or case ascertainment. A low proportion of smear-positive cases may result from poor laboratory techniques or excessive diagnosis of tuberculosis with reliance on chest X-rays and low interest in obtaining sputa for smears or cultures.

Once identified, tuberculosis cases are placed under treatment. Treatment of infectious cases is an important preventive measure. Treatment efficacy is evaluated by sputum conversion (both on smear and culture) of the active pulmonary cases. After 2 months of an effective regimen, 85% of active pulmonary cases should have converted their sputum from positive to negative. Therefore, the rate of sputum conversion at 2 months becomes an important indicator of program effectiveness. This indicator must be calculated for those who are smear positive and with a lesser importance for the other active pulmonary cases.

To ensure adequate treatment and prevent the development of acquired resistance, tuberculosis cases are placed under directly observed therapy (DOT). This measure is quite labor intensive. Priority must therefore be given to those at highest risk of relapse. These are the smear-positive culture-proven active pulmonary cases. DOT on extrapulmonary cases is much less important from a public health standpoint because they are not infectious and the major objectives of any public health program are to prevent transmission.

The same considerations apply to contact investigation and preventive treatment in countries that can afford a tuberculosis contact program. A recently infected contact is at the highest risk of developing tuberculosis the first year after infection; hence, the best preventive return is to identify contacts of infectious cases. Those contacts are likely to have been recently infected. Systematic screening of large population groups would also identify infected individuals, but most would be "old" infections at lower risk of developing disease. Individuals infected with tuberculosis and HIV are at extremely high risk of developing active tuberculosis. Therefore, the tuberculosis control program should focus on the population at high risk of HIV infection.

Often, program evaluation is performed by epidemiologists who have not taken the time to understand the dynamics of a disease in the community. Rates or proportions are calculated, no priorities are established, and precious resources are wasted on activities with little preventive value. For example, attempting to treat all tuberculosis cases, whether pulmonary or not with DOT, investigating all contacts regardless of the bacteriological status of the index case would be wasteful.

51.12 Mathematical Models

51.12.1 Aims of Mathematical Modeling

Mathematical models are an important tool for understanding the transmission dynamics of infectious diseases. In contrast to statistical models, dynamic transmission models are based on first principles. They aim at deriving population level phenomena from a mechanistic description of transmission between individuals of the population. The centerpiece of a dynamic transmission model consists of a term that quantifies the rate with which susceptible and infectious persons have contact with each other and the probability that transmission takes place during such a contact. In its simplest form, this term is modeled as a mass action term. In analogy to the mass action law in chemistry, the underlying assumption is that susceptible and infectious persons mix homogeneously and contact each other with a rate that is proportional to the concentrations of either population group. In more formal terms, if we denote by $X(t)$ the fraction of susceptible persons, and by $Y(t)$ the fraction of infected and infectious persons at time t, the rate at which infectious contacts take place is proportional to $X(t)^*Y(t)$. The proportionality factor β is a product of the contact rate – the number of contacts per unit time – and the probability that upon contact transmission of infection takes place. With the assumption that recovery from the infectious state into the immune state occurs with a constant rate γ, we can now formulate a first simple mathematical model:

$$\begin{aligned}\frac{dX(t)}{dt} &= -\beta X(t)Y(t),\\ \frac{dY(t)}{dt} &= \beta X(t)Y(t) - \gamma Y(t),\\ \frac{dZ(t)}{dt} &= \gamma Y(t),\end{aligned}$$

where $Z(t)$ is the fraction of immune or recovered individuals at time t. This set of equations is the simplest version of the so-called *susceptible-infected-removed* model that was first introduced in a more elaborate form by Kermack and McKendrick (1927; reprinted 1991). Since then, numerous variants of this simple model have been formulated and analyzed (Anderson and May 1991).

The primary aim of such a model is to gain a better understanding of the dynamics of the system. For example, in the above system of equations, we are interested in how an outbreak evolves in the population after introduction of a small number of index cases. The epidemic curve will depend on the parameters of the model, which in this case are the transmission rate β and the recovery rate γ. Let us see what we can say by just looking at the equations. Let us assume that we start at $t = 0$ in a situation where almost the entire population is susceptible, a small fraction of the population is infected, and nobody is immune. First, we observe that the fraction of susceptibles in the population can only decrease, because the right-hand side of the equation for $X(t)$ is negative. Furthermore, we see that $Y(t)$ will increase if the right-hand side of the equation for $Y(t)$ is positive, which is the case if $\beta X(0)Y(0) > \gamma Y(0)$. This leads to the insight that an outbreak is only possible if $\beta/\gamma > 1$. Here β/γ is the so-called basic reproduction number denoted by R_0 (see also Sect. 51.12.2). We will come back to this important concept later. Finally, we see that the fraction of immune persons in the population is continuously increasing as long as there are infected persons in the population. Figure 51.12 shows the typical time course of an outbreak for the parameters $\beta = 5$ and $\gamma = 1$.

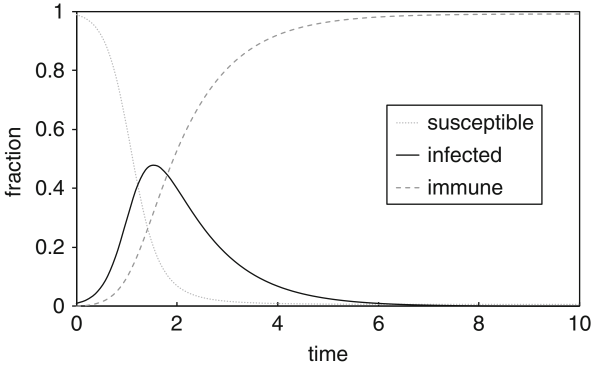


Fig. 51.12 The time course of an epidemic when there is no inflow of new susceptible individuals

After the outbreak has swept through the population and has left a certain fraction of the population immune, no further dynamics are possible in this system. This can only change if new susceptible individuals enter the population.

We want to extend the above model to include a simple demographic process, that is, births into the population with a rate ν and a per capita mortality rate μ. We assume that the mortality is not disease related, but applies to all population groups in the same way. For populations that grow or decrease in size, there are different ways of taking population size into account in the model formulation (Keeling and Rohani 2008). For simplicity, we assume here that $\nu = \mu$, meaning that the population is neither growing nor decreasing in size. With these additions, our model equations can be written as

$$\begin{aligned}\frac{dX(t)}{dt} &= \nu - \beta X(t)Y(t) - \mu X(t),\\ \frac{dY(t)}{dt} &= \beta X(t)Y(t) - \gamma Y(t) - \mu Y(t),\\ \frac{dZ(t)}{dt} &= \gamma Y(t) - \mu Z(t).\end{aligned}$$

These additions to the model change the long-term dynamic behavior of the model completely. The demographic process allows for a flow of new susceptible individuals into the population, thereby providing fuel for the transmission process to continue. In the long run, if the transmission rate β is large enough, the system will settle down to an endemic steady state with a prevalence of infection that is constant in time. It is not difficult to compute the endemic prevalence as a function of the model parameters. The result in terms of fractions X, Y, and Z of the population is

$$\begin{aligned}X &= \frac{1}{R_0},\\ Y &= \left(1 - \frac{1}{R_0}\right)\frac{\mu}{\gamma + \mu},\\ Z &= 1 - X - Y.\end{aligned}$$

In other words, the fraction of susceptible individuals is completely determined by the basic reproduction number R_0, with higher values of R_0 leading to a smaller proportion of susceptible individuals in the population. The endemic prevalence Y increases with increasing R_0 but is also determined by the duration of the infection $1/(\gamma + \mu)$.

Whether to use a model with or without demographic parameters depends on the time scale on which the outbreak takes place. For example, on the time scale of an influenza outbreak that takes a few weeks, the demographic process in the population will hardly influence the shape of the outbreak, and we might also not be interested in what will happen after the first wave. However, for an infection like

HIV, transmission dynamics evolves on the time scale of decades and will therefore strongly interact with the demographic process in the population.

This short introduction into the susceptible-infected-removed (SIR) model shows that (a) the model can provide insight into qualitative features of the transmission process on population level and (b) the exact form of the model to be used depends on the time scale and properties of the specific infection.

51.12.2 Important Concepts

Insights into the dynamics of the SIR model have led to the definition of a number of important concepts that are universal for all models and all infectious diseases. The most important of these concepts is the *basic reproduction number* R_0. The basic reproduction number R_0 is the number of secondary cases caused by one index case during his/her entire infectious period in a susceptible population. In other words, the basic reproduction number is given by the product of the transmission rate (number of new infections per time unit) and the duration of the infectious period. In the above model, the number of new infections per unit time is given by β, while the duration of the infectious period can be computed as $1/(\gamma + \mu)$. Therefore, for the SIR model with demographic process, we get

$$R_0 = \frac{\beta}{\gamma + \mu}.$$

We easily see from the equation describing the dynamics of $Y(t)$ that $Y(t)$ will increase in size if $R_0 > 1$ and decrease otherwise. This is the so-called threshold property that was already in 1927 formulated by Kermack and McKendrick (reprinted 1991) in a more general form. If $R_0 > 1$ an infected individual replaces himself by more than 1 new infected persons, which leads to an expansion of the epidemic. If $R_0 < 1$ the infection cannot establish itself in the population and will die out.

In more generality, one talks about the reproduction number $R(t)$ that describes the number of secondary cases per infected individual in a population that is not necessarily completely susceptible. $R(t)$ depends on the time, because with unfolding of the epidemic outbreak, the proportion of susceptibles in the population decreases and the proportion of immune individuals increases. Therefore, $R(t)$ will decrease with time. If there is no replenishment of the susceptible population, the transmission will stop at some point. The total number or fraction of the population that was infected during the entire course of the outbreak is called the *final size* of the epidemic. In epidemiological terms, the final size is called the attack rate; we denote it by A. The final size can be expressed in terms of the basic reproduction number as $A = 1 - \exp(-R_0 A)$ (Fig. 51.13).

If there is replenishment of the susceptible population by birth or recruitment into the population, the reproduction number may eventually converge to 1 in the endemic steady state. The prevalence in endemic steady state depends on the basic reproduction number with a larger R_0 leading to higher prevalence in steady state.

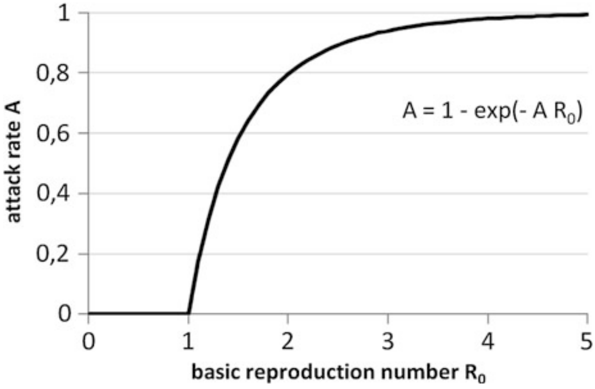


Fig. 51.13 The attack rate A as a function of the basic reproduction number R_0

Incorporating a simple term for infant vaccination with coverage p into the model makes it possible to determine the *critical vaccination coverage*. A model with vaccination is given by

$$\frac{dX(t)}{dt} = \nu(1-p) - \beta X(t)Y(t) - \mu X(t),$$
$$\frac{dY(t)}{dt} = \beta X(t)Y(t) - \gamma Y(t) - \mu Y(t),$$
$$\frac{dZ(t)}{dt} = \nu p + \gamma Y(t) - \mu Z(t),$$

where a fraction p of all newborns enters into the immune compartment immediately at birth, while the remaining fraction $1-p$ remains susceptible. Based on the endemic prevalence for this system of equations, the critical vaccination coverage can be derived as

$$p_{\text{crit}} = 1 - \frac{1}{R_0}.$$

This relationship gives valuable information about the vaccination effort needed to eliminate an infection from a population. It explains why it has been possible to eradicate smallpox with an estimated R_0 of around 5 (Gani and Leach 2001) in contrast to measles, for which the R_0 is estimated at around 20 (Wallinga et al. 2003).

Since the introduction of the SIR model, many different models have been formulated, and the diversity of models has vastly increased. Models have been designed for many specific infectious diseases, and they have incorporated population structure such as age, gender, spatial distribution, and differences in risk behavior. Also, stochastic models have been used to account for effects of chance events.

When choosing a model to answer a particular question in epidemiology or public health, different aims can be achieved, and the model of choice has to be accommodated to the aim. For answering questions about the qualitative dynamics of an infection, it is preferable to turn to a relatively simple model for which mathematical analysis is possible, while for the purpose of generating quantitative estimates or projections, more complex models are necessary that incorporate more details of the population structure (e.g., age) and more details about the transmission and course of infection. The latter can be simulation models that are implemented as computer code and cannot be formulated in terms of mathematical equations.

51.12.3 Use of Mathematical Models in Epidemiological Studies

Although mathematical modeling has been around for a long time, until recently, it was not much used as a tool for public health, but was considered a specialized research area for applied mathematicians and theoretical biologists. This started to change with the advent of the HIV pandemic, when mathematical models were first used to predict future epidemic spread, and to analyze the impact of behavior change on HIV incidence (Kaplan and Brandeau 1994). However, the breakthrough for mathematical modeling as a public health tool came with the concerns that smallpox virus could be used in a deliberate release and lead to devastating outbreaks in the only partially immune populations of present societies. How can public health policy be developed against threats with pathogens that are not circulating at present? There is no way to conduct epidemiological investigations, and the only available data in the case of smallpox were from before the eradication era. Therefore, to design policy, knowledge from historical smallpox outbreaks had to be combined with data about present-day society, and possible interventions had to be tested on the basis of this available information. Mathematical modeling provided a flexible tool to do that and was used to analyze possible vaccination strategies and other interventions (Ferguson et al. 2003).

Later, the experience with the global spread of SARS – the severe acute respiratory syndrome caused by a novel strain of corona virus – and the threat of a future pandemic with a new strain of influenza A initiated national and international efforts to better prepare for large outbreaks of emerging infections. Mathematical modeling was widely used for investigating optimal strategies for dealing with a new influenza pandemic (Longini et al. 2004; Ferguson et al. 2006). These response plans came into action during the pandemic with new influenza A|H1N1 emanating from Mexico in the spring of 2009. Even as the pandemic was still unfolding, first mathematical modeling studies started to deliver valuable data analyses almost in real time (Fraser et al. 2009).

Besides supporting public health policy in designing prevention and intervention strategies, mathematical modeling of infectious diseases has contributed greatly to increasing the understanding of the intricate relationships between clinical and biological determinants of infection and human contact and risk behavior patterns that lead to transmission. The importance of core groups of high sexual activity in the transmission dynamics of sexually transmitted infections (Hethcote and Yorke 1984), the impact of concurrent partnerships on the spread of HIV (Morris and Kretzschmar 1997), the importance of hosts being infectious before the appearance of symptoms for disease control (Fraser et al. 2004), and the connectedness of modern societies in a small world network (Watts and Strogatz 1998) are just some examples for how mathematical modeling has shaped the present paradigms of infectious disease epidemiology.

51.13 Conclusions

Today, the world is smaller than ever before, and international travel and a worldwide food market make us all potentially vulnerable to infectious diseases no matter where we live.

New pathogens are emerging such as the SARS or spreading through new territories such as WNV. WNV introduced in the USA in 1999 became endemic in the USA over the next years. Hospital-associated and community-associated methicillin-resistant *Staphylococcus aureus* (MRSA) and resistant tuberculosis cases and outbreaks are on the rise. Public health professionals are concerned that a novel recombinant strain of influenza will cause a new pandemic.

But not only the world and the etiological agents are changing; the world population is changing as well. In industrialized countries, the life expectancy is increasing, and the elderly are more likely to acquire a chronic disease, cancer, or diabetes in their lifetime. Because of underlying conditions or the treatment of these diseases, older populations also have an increased susceptibility for infectious diseases and are more likely to develop life-threatening complications.

Knowledge in the field of infectious disease epidemiology is expanding. While basic epidemiological methods and principles still apply today, improved laboratory diagnoses and techniques help to confirm cases faster, see how cases are related to each other, and therefore can support the prevention of spread of the specific disease. Better computers can improve the data analysis, and the Internet allows access to in-depth disease-specific information. Computer connectivity improves disease reporting for surveillance purposes, and the epidemiologist can implement faster preventive measures if necessary and is also able to identify disease clusters and outbreaks on a timelier basis.

The global threat of bioterrorism adds a new dimension. The intentional release of anthrax spores and the infection and death of persons who contracted the disease created a scare of contaminated letters in the US population.

With all these changes, there is renewed emphasis on infectious disease epidemiology and makes it a challenging field to work in.

References

Aintablian N, Walpita P, Sawyer MH (1998) Detection of Bordetella pertussis and Respiratory syncytial virus in air samples from hospital rooms. Infect Control Hosp Epidemiol 19:918–923

Alland D, Kalkut GE, Moss AR, McAdam RA, Hahn JA, Bosworth W, Drucker E, Bloom BR (1994) Transmission of tuberculosis in New York City. An analysis by DNA fingerprinting and conventional epidemiologic methods. NEJM 330(24):1710–1716

American Academy of Pediatrics (AAP) (2006) In: Pickering LK (ed) Red book: 2006 report of the committee on infectious diseases, 27th edn. AAP, Elk Grove Village

American Public Health Association (2008) In: Heymann DL (ed) Control of communicable diseases manual, 19th edn. United Book Press Inc, Baltimore

Anderson RM, May RM (1991) Infectious diseases in humans: transmission and control. Oxford University Press, Oxford

Balsamo G, Michaels S, Sokol T, Lees K, Mehta M, Straif-Bourgeois S, Hall S, Krishna N, Talati G, Ratard R (2003) West Nile epidemic in Louisiana in 2002. Ochsner Health J 5(3):13–15

Boelaert M, Arbyn M, Van der Stuyft P (1998) Geographical Information System (GIS), gimmick or tool for health district management? Trop Med Int Health 3(3):163–165

Brachman PS (1998) Epidemiology of nosocomial infections, Chapter 1:3. In: Bennett JV, Brachman PS (eds) Hospital infections, 4th edn. Lippincott-Raven Publishers, Philadelphia

Buehler JW, Hopkins RS, Overhage JM, Sosin DM, Tong V (2004) Framework for evaluating public health surveillance systems for early detection of outbreaks. Recommendations of a CDC Working Group, MMWR 53 (RR05) May 7

CDC (1985) Update: International outbreak of restaurant-associated botulism – Vancouver, British Columbia, Canada. MMWR 34(41):643

CDC (1987) Epidemiologic notes and reports restaurant associated botulism from mushrooms bottled in-house – Vancouver, British Columbia, Canada. MMWR 36(7):103

CDC (1996) Outbreaks of cyclospora cayetanensis infection – United States, 1996. MMWR 45(25):549–551

CDC (1997) Case definitions for infectious conditions under public health surveillance. MMWR 46(10):1–55

CDC (1999) Framework for program evaluation in public health. MMWR 48(11):1–40

CDC (2001a) "Norwalk-like viruses": public health consequences and outbreak management. MMWR 50(9):1–17

CDC (2001b) Updated guidelines for evaluating public health surveillance systems: recommendations from the guidelines working group. MMWR 50(13):1–35

CDC (2002) Outbreaks of gastroenteritis associated with Noro viruses on cruise ships – United States, 2002. MMWR 51(49):1112–1115

CDC (2004) Diagnosis and management of foodborne illnesses: a primer for physicians and other health care professionals. MMWR 53(4):1–33

CDC (2006) National antimicrobial resistance monitoring system for enteric bacteria (NARMS): 2003 Human Isolates Final Report. U.S. Department of Health and Human Services, Atlanta, Georgia

CDC (2011) CDC estimates of foodborne illness in the US. http://www.cdc.gov/foodborneburden/PDFs/FACTSHEET_A_FINDINGS_updated4-13.pdf/. Accessed 3 Jan 2012

Ferguson NM, Keeling MJ, Edmunds WJ, Gani R, Grenfell BT, Anderson RM, Leach S (2003) Planning for smallpox outbreaks. Nature 425(6959):681–685

Ferguson NM, Cummings DA, Fraser C, Cajka JC, Cooley PC, Burke DS (2006) Strategies for mitigating an influenza pandemic. Nature 442(7101):448–452

Fraser C, Riley S, Anderson RM, Ferguson NM (2004) Factors that make an infectious disease outbreak controllable. Proc Natl Acad Sci USA 101(16):6146–6151

Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, Griffin J, Baggaley RF, Jenkins HE, Lyons EJ, Jombart T, Hinsley WR, Grassly NC, Balloux F, Ghani AC, Ferguson NM, Rambaut A, Pybus OG, Lopez-Gatell H, Alpuche-Aranda CM, Chapela IB, Zavala EP, Guevara DM, Checchi F, Garcia E, Hugonnet S, Roth C, WHO Rapid Pandemic Assessment Collaboration (2009) Pandemic potential of a strain of influenza A (H1N1): early findings. Science 324(5934):1557–1561

Frost WH (1936) Snow on cholera. In: Frost WH (ed) Harvard University Press, Cambridge, MA

Gani R, Leach S (2001) Transmission potential of smallpox in contemporary populations. Nature 414(6865):748–751. Erratum in: Nature 2002, 415(6875):1056

Goodwin LG, Gordon Smith CE (1996) Yellow fever. In: Cox FEG (ed) The Wellcome Trust illustrated history of tropical diseases. Wellcome Trust, London, p 147

Haber M, Barskey A, Baughman W, Barker L, Whitney CG, Shaw KM, Orenstein W, Stephens DS (2007) Herd immunity and pneumococcal conjugate vaccine: a quantitative model. Vaccine 25(29):5390–5398

Hennessy TW, Hedberg CW, Slutsker L, White KE, Besser-Wiek JM, Moen ME, Feldman J, Coleman WW, Edmonson LM, MacDonald KL, Osterholm MT (1996) A national outbreak of salmonella enteritidis infections from ice cream. NEJM 334(20):1281–1286

Hethcote HW, Yorke JA (1984) Gonorrhea transmission dynamics and control. Springer, New York

Hinds MW, Skoggs JW, Bergeisen GH (1985) Benefit-cost analysis of active surveillance of primary care physicians for Hepatitis A. Am J Public Health 75:176–177

Hornick R, Greisman SE, Woodward TE, DuPont HL, Drawkins AT, Snyder MJ (1970) Typhoid fever: pathogenesis and immunologic control. NEJM 283:686–691

Jones SC, Morris J, Hill G, Alderman M, Ratard RC (2002) St. Louis encephalitis outbreak in Louisiana in 2001. La State Med Soc 154(6):303–306

Kaplan EH, Brandeau ML (eds) (1994) Modelling the AIDS epidemic: planning, policy and prediction. Raven Press, New York

Keeling MJ, Rohani P (2008) Modeling infectious diseases in humans and animals. Princeton University Press, Princeton/Oxford

Kermack WO, McKendrick AG (1991) Contributions to the mathematical theory of epidemics – I. 1927. Bull Math Biol 53(1–2):33–55

Lengeler C, Makwala J, Ngimbi D, Utzinger J (2000) Simple school questionnaires can map both Schistosoma mansoni and Schistosoma haematobium in the Democratic Republic of Congo. Acta Trop 74(1):77–87

Lockman S, Sheppard JD, Braden CR (2001) Molecular and conventional epidemiology of Mycobacterium tuberculosis in Botswana: a population-based prospective study of 301 pulmonary tuberculosis patients. J Clin Microbiol 39(3):1042–1047

Longini IM Jr, Halloran ME, Nizam A, Yang Y (2004) Containing pandemic influenza with antiviral agents. Am J Epidemiol 159(7):623–633

Loudon RG, Bumgarner LR, Lacy J, Coffman GK (1969) Aerial transmission of mycobacteria. Am Rev Respir Dis 100(2):165–171

Mandell GL, Douglas R, Bennett JE (2000) Principles and practice of infectious diseases, 5th edn. Churchill Livingstone, Philadelphia

McCullough NB, Eisele CW (1951) Experimental human salmonellosis II. Pathogenicity of strains of salmonella newport, salmonella derby, and salmonella bareilly obtained from spray-dried eggs. J Infect Dis 89:209–213

Merlos I (2002) An uninvited guest at "Turkey Day." Louisiana Morb Rep 13(1):1–2

Michaels SR, Balsamo GA, Kukreja M, Anderson C, Straif-Bourgeois S, Talati G, Ratard RC (2005) Surveillance for West Nile Virus cases in Louisiana 2001–2004. J La State Med Soc (157):269–272

Morris M, Kretzschmar M (1997) Concurrent partnerships and the spread of HIV. AIDS 11:641–648

Mounts AW, Holman RC, Clarke MJ, Bresee JS, Glass RI (1999) Trends in hospitalizations associated with gastroenteritis among adults in the United States, 1979–1995. Epidemiol Infect 123:1–8

Porta M, Greenland S, Last J (2008) A dictionary of epidemiology, 5th edn. Oxford University Press, Oxford

Riley RL, Mills CC, Nyka W, Weinstock N, Storey PB, Sultan LU, Riley MC, Wells WF (1956) Aerial dissemination of pulmonary tuberculosis: a two-year study of contagion in a tuberculosis ward. Am J Hyg 70:185–196 (Reprinted in the Am J Epidemiol 1995, 142:1–14)

Schulman JL (1970) Transmissibility as a separate genetic attribute of Influenza viruses. In: Silver IH (ed) Aerobiology. Academic Press, New York, p 248

Snacken R, Kendal AP, Haaheim LR, Wood JM (1999) The next influenza pandemic: lessons from Hong Kong, 1997. Emerg Infect Dis 5(2):195–203

Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV, The CDC PulseNet Task Force (2001) PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerg Infect Dis 7(3):382–389

Toman K (2004) Toman's tuberculosis case detection, treatment and monitoring: questions & answers, 2nd edn. World Health Organization, Geneva

University of California (2011) Unequal causes of death. http://ucatlas.ucsc.edu/. Accessed 16 Dec 2011

Van Beneden CA, Lexau C, Baughman W, Barnes B, Bennett N, Cassidy PM, Pass M, Gelling L, Barrett NL, Zell ER, Whitney CG (2003) Aggregated antibiograms and monitoring of drug-resistant *Streptococcus pneumoniae*. Emerg Infect Dis. http://www.cdc.gov/ncidod/EID/vol9no9/02-0620.htm. Accessed 16 Dec 2011

Wallinga J, Teunis P, Kretzschmar M (2003) Reconstruction of measles dynamics in a vaccinated population. Vaccine 21(19–20):2643–2650

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442

WHO (2008) Infectious disease report, Chapter 2. http://www.who.int/infectious-disease-report/pages/ch2text.html. Accessed 16 Dec 2011

WHO (2010) Fact sheets on infectious diseases. http://www.who.int/mediacentre/factsheets/en/. Accessed 16 Dec 2011

Wills C (1996a) Cholera, the black one. In: Yellow fever black goddess, the coevolution of people and plagues. Addison-Wesley Publishers, Reading, p 115

Wills C (1996b) Four tales of the new decameron. In: Yellow fever black goddess, the coevolution of people and plagues. Addison-Wesley Publishers, Reading, p 84

Cardiovascular Health and Disease 52

Darwin R. Labarthe

Contents

D.R. Labarthe
Department of Preventive Medicine, Northwestern University Feinberg School of Medicine,
Chicago, IL, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_35,

52.1 Introduction

Worldwide, coronary heart disease (CHD) and stroke are the first and second leading causes of death, major contributors to disability, and the most common conditions – following hypertension or high blood pressure – comprising the cardiovascular diseases (CVDs) (Mazzati et al. 2006). Together these conditions rank among the foremost public health challenges of our time. They have been under extensive epidemiological investigation for more than a half century. As a result, with parallel clinical and laboratory research, understanding of the causation and means of prevention of CVDs have become well established. Their present and future global impact is becoming recognized. Calls for action to prevent heart disease and stroke are widespread at national and global levels (Labarthe 2011).

Three resources of potential interest to readers are (1) a comprehensive public health textbook on cardiovascular epidemiology and prevention (Labarthe 2011); (2) an extensive historical archive of materials regarding people, studies, essays, oral history, and audiovisuals from the field of cardiovascular epidemiology, found on this dedicated website: www.epi.umn.edu/cvdepi/ (Blackburn 2012); and (3) an annotated bibliography of some 150 classic and contemporary publications on public health aspects of heart disease (Labarthe 2013).

This chapter begins with a discussion of the scope and basic concepts of cardiovascular epidemiology. As the main focus of this chapter, the atherosclerotic and hypertensive diseases are defined, the main types of studies are illustrated, and a public health perspective on cardiovascular health and disease is presented. Against this backdrop, the brief sections that follow address these aspects of cardiovascular epidemiology:

- An epidemiological description
- Determinants
- Causation and prevention
- Research needs

52.2 Scope and Basic Concepts

52.2.1 Atherosclerotic and Hypertensive Diseases

CHD and stroke are the main consequences of atherosclerosis and hypertension. Also termed ischemic heart disease (IHD), CHD is a manifestation of reduced blood supply – and hence reduced oxygen supply – via the coronary arteries to the myocardium (muscle of the heart). Sudden loss of oxygen supply may result in injury or death to the muscle cells, constituting a myocardial infarction (heart attack). Disturbance of normal rhythmic contraction of the heart may occur, and this arrhythmia often causes sudden death unless effective emergency aid is administered promptly. A less severe – but also ominous, perhaps equally painful, and recurring – condition is unstable angina, which represents transient impairment of blood flow. The typical culprit lesion in the artery wall is the atherosclerotic plaque, which may itself grow gradually over years to narrow or occlude the artery or may suddenly rupture and trigger the rapid formation of an occlusive thrombus (blood clot) within the artery. Survival from an acute coronary event may be accompanied by disability from residual cardiac impairment and entails a high risk of recurrent cardiovascular events or progressive heart failure.

Stroke results from injury to the brain in a manner sometimes analogous to myocardial infarction: interruption of the arterial blood supply by a thrombotic occlusion. But other mechanisms may also cause stroke, chiefly occlusion by lodging of an embolus (a blood clot that has arisen, e.g., in a damaged heart and has been carried through the circulation to the brain) or by hemorrhage (when a blood vessel within the brain ruptures). Regardless of the mechanism, interruption of blood flow may produce injury and death of brain cells. The event may be rapidly fatal or may be followed by survival, with a high risk of recurrence and often with significant disability. The main underlying condition leading to thrombosis is atherosclerosis, whereas hemorrhage is especially attributable to hypertension.

Chronic heart failure may develop as a consequence of either CHD or hypertension, greatly increases the risk of stroke, and causes disability due to impaired circulatory function. It may also result from several other cardiac disorders, but in populations where CHD is a frequent condition, heart failure as a late consequence of CHD has become increasingly common, especially among older adults.

These five interrelated conditions – atherosclerosis and hypertension and consequent CHD, stroke, and heart failure – comprise the greater part of CVD, as used throughout this chapter. Other important cardiovascular conditions, such as atherosclerotic peripheral arterial disease, aortic aneurysm, cardiomyopathies, rheumatic heart disease, Chagas' disease, congenital heart disease, deep vein thrombosis, and pulmonary embolism, while important vascular conditions, are beyond the scope of this chapter.

A recent development in the global health arena, commonly reflected in national-level discussions as well, is a focus on several chronic diseases together, rather than separately. Thus, "non-communicable diseases (NCDs)" include CVDs, cancer, chronic respiratory diseases, and diabetes (World Health Organization 2011; Bloom et al. 2011). These discussions are directly relevant to CVDs, even when they are not addressed explicitly, because they are a major, often dominant, component of the NCDs. This explains reference to the NCDs at the close of the chapter, in relation to policy and research in the field of cardiovascular health.

52.2.2 Types of Studies

Measures of CVD in the population are fundamental to epidemiological investigation and understanding. Terms such as mortality, incidence, or prevalence introduced, for example, in Part I (Concepts and Designs in Epidemiology) of this handbook have their particular use and importance in CVD epidemiology. *CVD mortality* can be studied insofar as deaths are registered and classified in accordance with reliable and standardized procedures and the current census of the underlying population is known. Such data have been collected for several decades in many countries but have yet to be recorded at all in many others. The most informative data are those presented for specific age and sex groups within a population and for other demographic subgroups as appropriate. Adherence to practices for the *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision* (World Health Organization 1992) is now required for national death registration systems. (The next revision, the 11th, is scheduled for release in 2015 – see World Health Organization (2012).) For epidemiological research, cause-of-death validation can be conducted through well-standardized procedures such as those of the WHO MONICA Project (World Health Organization MONICA Project Principal Investigators 1988). Validation is especially important for unobserved and out-of-hospital deaths in which CHD or stroke is suspected, but documentation regarding diagnostic criteria is unavoidably incomplete.

The CVD burden of a population may be estimated, and comparisons among populations made, by surveys to determine the proportion of the population affected, or *CVD prevalence*. Prevalence of CVD is influenced by both the rate at which new cases occur and the duration of their survival in the population. Sampling methods and selective factors in survey participation, such as disease-related disability that prevents participation (and therefore precludes complete and unbiased representation of the study population), must be taken into account in evaluating findings. Standard cardiovascular survey methods have been published since the 1960s (Blackburn et al. 1960; Rose and Blackburn 1968; Luepker et al. 2004). The prevalence of CVD and related conditions such as high blood pressure, high blood cholesterol concentration, and smoking history, as well as such major underlying factors as diet and physical activity, can be assessed with reasonable efficiency.

The frequency with which newly detected cases of CVD occur in a population, or *CVD incidence*, is measured by long-term (several years) follow-up among persons found at a baseline survey to be disease-free and for whom first events can be ascertained by periodic reexamination, continuous surveillance of hospitalizations or deaths, or a combination of these approaches. Issues in evaluation of findings relate especially to chances for missed or misdiagnosed cases and losses to follow up. Detection of new cases also allows assessment of the proportion of events leading to death in the short-term (*case fatality*; usually within 28 days of symptom onset) and *long-term survival* (for any specified period beyond 28 days post-event).

The now familiar types of epidemiological studies of chronic diseases were devised or refined over the decades since the 1950s, often through research on CVDs. They can be described as follows (see also Part I (Concepts and Designs in Epidemiology) of this handbook):

- *Analysis of vital statistics*, specifically comparison of death rates (mortality) from cardiovascular causes, was an early strategy in CVD epidemiology. It continues to be of value in demonstrating differences between population groups, or trends over time, in national or other geopolitical areas. *Examples*: Gordon (1957) recognized that men of Japanese ancestry living in Japan, Hawaii, and California experienced strikingly different patterns of mortality from CHD and stroke (CHD being low in Japan, intermediate in Hawaii, and high in California, in contrast to the opposite gradient for stroke). This analysis stimulated more intensive investigation through other epidemiological methods (Kagan and Yano 1996). Thom and others (1992) showed marked short-term variation and revealing patterns in CHD and stroke death rates in 27 countries over four decades. A major strength of this approach is ready availability of vital data over decades or longer for many countries and their subdivisions. Serious limitations include the lack of data especially for many low- and middle-income countries and the need to take into account changes in classification of causes of death over time and data reliability.

- *Population surveys* (or "cross-sectional studies") are conducted in principle at a point in time (ongoing, continuous surveys are better regarded as *population surveillance*) and typically in a defined geographical area. In their simplest design, they can provide essential information about the proportion of a population affected by a particular disease (disease prevalence) and the population distribution of related factors. More extensive surveys can include multiple-population comparisons or repetition at intervals of years, thereby addressing population differences or secular trends. Prediction of findings permits testing of prior hypotheses, as well. *Examples*: The INTERSALT Study demonstrated an association between the slope of increasing blood pressure with age and urinary sodium and potassium excretion in adults among 52 study centers in 32 countries (INTERSALT Co-operative Research Group 1986). The Kenyan Luo Migration Study showed rapid change in blood pressure, over only weeks to months, among Luo tribesmen migrating from rural to urban Kenya (Poulter et al. 1990). From serial surveys representative of the United States population, favorable trends were seen in the distribution of systolic blood pressure over 40 years, from the 1960s to the mid-1990s (Goff et al. 2001). Comparison of national survey data on blood pressure revealed higher prevalence of hypertension and lower prevalence of hypertension control, in six European countries than in Canada and the United States (Wolf-Maier et al. 2003). Strengths of the population survey include versatility of design, fundamental simplicity, and relative expediency of execution. Limitations include the cross-sectional property that necessarily restricts observation to prevalent cases and constrains collection of historical data on exposures that may have preceded onset of disease.
- *Cohort studies* are designed to determine the rate of occurrence of new cases among persons free of disease at first, or "baseline" observation and to measure disease incidence in relation to baseline characteristics of interest. Cohort studies were initiated in the late 1940s and after to yield insights into causes of CHD and stroke. *Examples*: The Framingham Heart Study, begun in 1948, now includes some 60 years of multigenerational observation of CVD in one US community (Dawber et al. 1951). The Whitehall Study of British civil servants also provided long-term follow-up and insight to individual risks (Rose and Shipley 1986). The Seven Countries Study demonstrated major factors accounting for population differences in CHD rates between Northern and Southern Europe, Japan, and North America (Keys 1980). The several ongoing, long-term cardiovascular cohort studies in the USA are described periodically in government publications

(National Heart, Lung, and Blood Institute 2006). Cohort studies are unique among observational designs in permitting direct assessment of disease incidence in relation to baseline or interim risk characteristics, or relative risk. In the CVD arena, they require large populations, long-term follow-up, or both to generate sufficient person-years of experience to afford reliable estimates of incidence and risk and are correspondingly demanding of resources and skill.

- The *case-control study* is used to identify already affected persons (the case group) in order to assess characteristics of interest among them; comparable assessment is then conducted among a suitable non-affected group (the control group). Comparative analysis permits judgment whether an exposure is sufficiently greater (or less) among cases and therefore may be linked in a meaningful way with the presence or absence of disease. *Examples*: The case-control approach has been applied in large-scale multinational collaborative studies. The association between the presence of stroke and other vascular conditions among women by their history of oral contraceptive use was investigated in the WHO Collaborative Study of Cardiovascular Disease and Steroid Hormone Contraception (World Health Organization 1996). More recently several factors of interest were studied in relation to the presence or absence of myocardial infarction among men and women in 52 countries throughout the world (Yusuf et al. 2004). This approach is widely used to study conditions of relatively low incidence, a situation in which demands of cohort studies noted above are difficult to satisfy. The fundamental issue in design, conduct, and interpretation of case-control studies is evaluation of possible bias or confounding in the comparison of cases and controls (see chapters ▶Confounding and Interaction and ▶Sensitivity Analysis and Bias Analysis of this handbook).
- *Combinations of approaches* have also been used to monitor CHD and stroke in populations. *Example*: An especially comprehensive study of this kind was the WHO MONICA Project, a major collaboration including 38 populations in 21 countries, chiefly in Europe (Monitoring Trends and Determinants in Cardiovascular Disease Project) (World Health Organization MONICA Project Principal Investigators 1988; Sarti et al. 2003; Tunstall-Pedoe 2003). Repeated *cross-sectional surveys* were used to assess trends in risk factors and treatment of cases, continuous *community-wide and hospital-based surveillance* identified new and recurring cases and case fatality, and *vital statistics* systems supported analysis of death rates based on validated cause of death in accordance with the MONICA Project surveillance protocol. In some centers, *cohort designs* were incorporated as well. Composite study designs combine both advantages and limitations of each approach that is included.
- The many "clues to causation" generated from the types of studies described here, characterized collectively as "observational studies," have led to a very large body of experimental research, such as the intervention trials addressed in chapters ▶Intervention Trials, ▶Cluster Randomized Trials, and ▶Community-Based Health Promotion of this handbook. In CVD epidemiology, this research includes *clinical trials* among selected individuals, to evaluate the efficacy of treatments among patients (*secondary prevention trials*) or of measures to reduce risk of first events (*primary prevention trials*). It also includes *community-based*

trials, and *demonstration projects* to test the effectiveness of interventions for the population at large (see, e.g., chapter ▶Community-Based Health Promotion of this handbook. *Examples*: Among early clinical trials that stimulated subsequent population-based trials and strongly influenced policy and practice is the US Veterans Administration trial of the treatment of hypertension (Freis 1990). Community trials and demonstrations are well illustrated by the North Karelia Project (Puska et al. 1998) and by later examples in the United States and elsewhere (Blackburn 1992; Labarthe 2011).

52.2.3 A Public Health Perspective

Extensive knowledge of the natural history of CVDs, and of interventions to modify their course, has accrued through application of all of the types of studies illustrated above. A synthesis of this knowledge from a public health perspective was undertaken a decade ago to provide a framework for public health action to prevent heart disease and stroke (Fig. 52.1) (US Department of Health and Human Services 2003). Concepts basic to CVD epidemiology, illustrated in Fig. 52.1, relate to the progression from cardiovascular health to disease and the corresponding array of intervention strategies and approaches that apply across this continuum. They are addressed briefly here as background for the sections that follow.

The lower panel of Fig. 52.1 represents as “the present reality” a well-established series of connections characterizing the progression from unfavorable social and

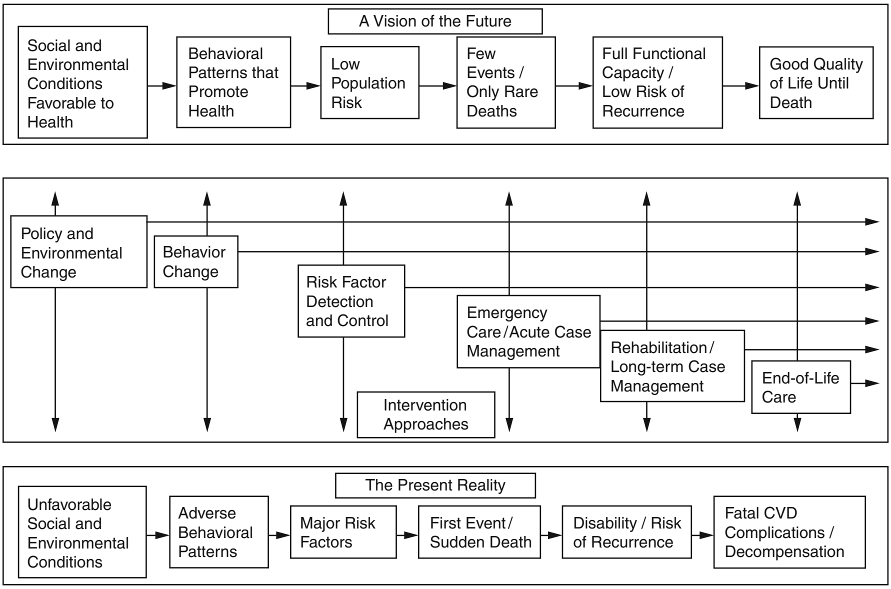

A Vision of the Future
Social and Environmental Conditions Favorable to Health
Behavioral Patterns that Promote Health
Low Population Risk
Few Events / Only Rare Deaths
Full Functional Capacity / Low Risk of Recurrence
Good Quality of Life Until Death
Policy and Environmental Change
Behavior Change
Risk Factor Detection and Control
Emergency Care / Acute Case Management
Rehabilitation / Long-term Case Management
End-of-Life Care
Intervention Approaches
The Present Reality
Unfavorable Social and Environmental Conditions
Adverse Behavioral Patterns
Major Risk Factors
First Event / Sudden Death
Disability / Risk of Recurrence
Fatal CVD Complications / Decompensation

Fig. 52.1 Action framework for a comprehensive public health strategy to prevent heart disease and stroke (Adapted from US Department of Health and Human Services 2003)

environmental conditions, through the adverse behavioral patterns that they foster, to the emergence of the major CVD risk factors. (Not shown, for simplicity, are the typical long-term development of subclinical, or undetected, disease and the immediate precipitating factors that link the major risk factors to the first event – whether heart attack, stroke, or episode of heart failure.) Sudden death may rapidly end the course, or survival may extend it, often with significant disability and high risk of recurrent CVD events. Paralleling this progression is the dimension of the life course of CVD, from possible maternal and fetal influences where earliest environmental factors may operate to childhood and adolescence where behavioral patterns often become established and risk factors begin to emerge and to early, middle, and late adulthood with the acute events and their consequences (see also chapter ▶Life Course Epidemiology of this handbook).

The counter to this present reality is a vision of the future (upper panel), in which the opposite of each of these states has been achieved. The means of change toward this vision are the intervention approaches identified in the center panel of this figure. Each approach has potential application from its first point of intervention throughout the further progression of CVD. For example, policy and environmental change can be applied to health-care settings, work sites, and schools as well as to society at large (Association of State and Territorial Directors of Health Promotion and Public Health Education 2001). Risk factor detection and control applies not only in advance of any clinical event but also, for persons who survive them, to lifelong care to prevent recurrent CVD events. These and the more familiar concepts of secondary, primary, and primordial prevention and of high-risk and population-wide strategies are discussed further below.

The epidemiological underpinnings of this framework derive from decades of research through the foregoing approaches. The public health implications of the framework and the plans of action that follow derive in turn from clinical and community trials, experience in public health practice, and knowledge gained through many years of policy development, implementation, and evaluation. The following picture of cardiovascular health and disease at the population level demonstrates the importance, for population health, of taking action on a large scale on the basis of this evidence.

52.3 An Epidemiological Description

52.3.1 Overview

More than a half century of CVD epidemiology, developing and applying the methods highlighted above, has provided a comprehensive understanding of the natural history of these conditions (see, e.g., Marmot and Elliott 2005; Labarthe 2011). Several broad features of this understanding are:

- CVD is pervasive throughout the world. In the mid-twentieth century, CVD, and especially CHD, came to be regarded as a public health problem of the Western industrialized nations. Before the end of the century, CVD (principally expressed as CHD or stroke) was recognized as a public health problem of global importance and a significant deterrent to social and economic development in low- and middle-income countries.

- Change in population burdens of CVD, as reflected in the crude, late measure of cause-specific mortality, can occur rapidly and with abrupt and unpredicted change in direction. Such rapid change is unequivocal evidence of powerful environmental factors whose impact is highly variable over time and place. In addition, migration of population subgroups has demonstrated change in CVD rates such that rates for migrants come to resemble those of the new population setting in contrast to rates for the non-migrating members of the source population. Such observations suggest that environmental factors are potentially subject to interventions to establish or maintain social and environmental conditions protective against CVD.
- Epidemiological investigation of the pathology of atherosclerosis strongly supports the view of early-life onset and lifelong progression of CVD. Studies of military casualties dying of non-CVD causes have shown that unrecognized coronary atherosclerosis may be extensive even in early adulthood. Thus, true prevention of atherosclerosis and its complications requires effective intervention as early as childhood and adolescence.
- National vital statistics systems that provide reliable data on cause-specific mortality are sometimes adequate for gauging the population burden of CVD, but these data are incomplete or lacking for many countries. Additional information is needed to fully characterize a population's CVD profile, and efforts to compile such data on a worldwide basis have revealed serious shortcomings even in high-income countries, such as the United States, where data on disease incidence are typically lacking except in specific localities with special epidemiological studies. Characterizing the global burden more definitively thus faces serious obstacles.
- Notwithstanding these limitations, serious efforts to project the global burden of CVD and other major causes of death and disability, and the potential impact of selected interventions against these conditions, have provided a view of the world's health two decades or more hence. The results indicate that CHD and stroke have been the first and second leading causes of death worldwide since 1990 and would remain so through 2020, barring greatly intensified public health efforts in prevention.

Selected examples below illustrate a twentieth-century historical perspective, the "geographical pathology" of CVD, and forecasts of the global CVD burden.

52.3.2 Illustrations

52.3.2.1 Historical Perspective

Studying long-term CVD mortality trends in one or more countries can provide valuable epidemiological insight. A leading example of this approach is the work of Omran (1971), who studied mortality from multiple causes in the United States

Table 52.1 The epidemiological transition (Source: Pearson et al. 1993)

Phase of epidemiological transition	Deaths from circulatory disease (%)	Circulatory problems	Risk factors
Age of pestilence and famine	5–10	Rheumatic heart disease; infectious and deficiency-induced cardiomyopathies	Uncontrolled infection; deficiency conditions
Age of receding pandemics	10–35	As above, plus hypertensive heart disease and hemorrhagic stroke	High-salt diet leading to hypertension; increased smoking
Age of degenerative and man-made diseases	35–55	All forms of stroke; ischemic heart disease	Atherosclerosis from fatty diets; sedentary lifestyle; smoking
Age of delayed degenerative diseases	Probably under 50	Stroke and ischemic heart disease[a]	Education and behavioral changes leading to lower levels of risk factors

[a] At older ages. Represents a smaller proportion of deaths

and elsewhere from 1900 through 1970. On the basis of these observations, he presented the concept of "epidemiological transition" to describe a relative shift in frequency between infectious and circulatory diseases as causes of death over long periods of cultural development. In an era where causes of death were dominated by "pestilence and famine," few deaths would be due to circulatory causes, that is, CVD. Successive eras of "receding pandemics" and rise of "degenerative and man-made diseases" would reach a peak in the proportion of deaths from CVD, followed by a decline brought about by "education and behavioral changes leading to lower levels of risk factors" (Table 52.1) (Pearson et al. 1993). This "theory" has become more a descriptor of historical trends than predictor of change, as it has been expanded to take more recent variations into account (Olshansky and Ault 1986; Yusuf et al. 2001). A further, serious limitation of the concept is its foundation in *proportionate* mortality: Where rates of infectious disease and maternal and infant mortality remain high, the proportion of CVD deaths may remain low, masking the true absolute increase in the CVD burden and delaying public health recognition of the urgency of this development.

A more recent analysis, noted above, focused on national secular trends in CHD and stroke mortality from the 1950s to the 1980s in 27 countries (Thom et al. 1992). For CHD, rates for men consistently exceeded those for women; rates ranged widely among countries, as much as six-fold; rates changed markedly within countries over successive 5-year intervals; and trends were generally similar for women and men in the same country (see Fig. 52.2). These observations indicate strong time-dependent and regionally or nationally distinct environmental influences that commonly, though not invariably, affected women and men in similar ways. The rapid pace of change excludes population genetics as an explanation.

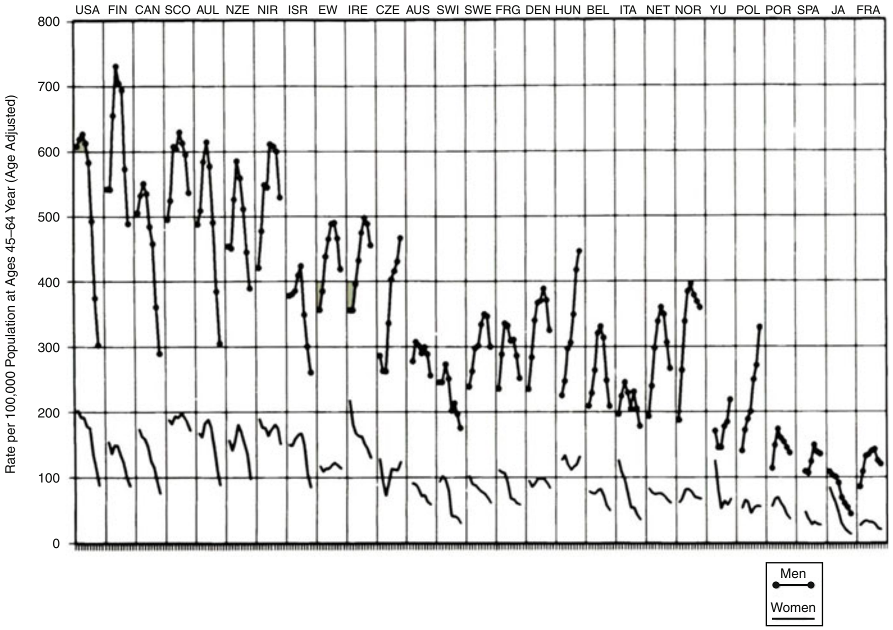


Fig. 52.2 Coronary heart disease mortality in 27 countries (USA, United States; FIN, Finland; CAN, Canada; SCO, Scotland; AUL, Australia; NZE, New Zealand; NIR, Northern Ireland; ISR, Israel; EW, England and Wales; IRE, Ireland; CZE, Czechoslovakia; AUS, Austria; SWI, Switzerland; SWE, Sweden; FRG, Federal Republic of Germany; DEN, Denmark; HUN, Hungary; BEL, Belgium; ITA, Italy; NET, Netherlands; NOR, Norway; YU, Yugoslavia; POL, Poland; POR, Portugal; SPA, Spain; JA, Japan; FRA, France. Each point represents the sex-specific average annual mortality for one period. Countries are arrayed in descending order of mortality for men in the first interval), 1950–1987 (Source: Thom et al. 1992)

For stroke, the same 27 countries appeared in quite different rank order (e.g., Japan was highest for stroke, though next to lowest for CHD), rates for men and women were generally similar within each country, and overall rates and trends were quite homogeneous among countries, excepting mainly several central European countries. Stroke does not closely match CHD epidemiologically. It does, however, share the aspect of marked change in mortality in a rather short historical period that indicates strong environmental influences, here (vs. CHD) operating in a generally similar way between sexes and among countries.

52.3.2.2 "Geographical Pathology" of Atherosclerosis

The term "geographical pathology" was coined and applied to atherosclerosis at least as early as 1934, being discussed in a conference with this title in the Netherlands in that year (Epstein 2005). Epidemiological study of atherosclerosis has included a substantial body of work by pathologists that is well represented by three studies – one in the true vein of geographical variation and the others each uniquely designed for study of postmortem epidemiology. The International

Atherosclerosis Project was essentially a cross-sectional multinational survey of coronary and aortic atherosclerosis (Tejada et al. 1968). Sections of coronary arteries and aortas were obtained from more than 23,000 autopsies for non-CVD deaths at ages 10–69 years in 14 countries, including Latin America as well as the United States and South Africa. Wide variation among the populations was found in the average extent of coronary atherosclerosis at any given age, and this measure and the CHD mortality rates corresponded closely within countries. Males had more extensive lesions than females did in countries where the extent of coronary atherosclerosis was greatest.

A strong age gradient in degree of atherosclerosis at death was also observed in six of the countries studied: From a range of 0–25% of intimal surface involvement with fibrous plaques at age 20, the percentage approximately doubled by age 30 in each population and continued to increase, although less steeply, to later ages at death. These findings confirmed marked population differences in the extent of coronary atherosclerosis, especially at early ages. They also supported earlier observations indicating that onset of atherosclerosis was common before age 20 and that males exhibited more extensive coronary atherosclerosis than females did (Holman et al. 1958).

The second example is the Pathobiological Determinants of Atherosclerosis in Youth (PDAY) Study, not geographical in orientation but aimed at the early development of atherosclerosis (Pathobiological Determinants of Atherosclerosis in Youth (PDAY) Research Group 1990). PDAY conducted standardized postmortem examinations of coronary and aortic specimens among black or white decedents aged 15–34 years from non-CVD causes at eight centers in the United States. Beyond pathology alone, this study introduced assessment of CVD risk factors based on blood samples and other materials obtained at death. The extent and severity of atherosclerosis at these early ages was strongly related to adverse blood lipid profiles and to smoking (as assessed by serum thiocyanate concentration). Earlier, as a third example, the Bogalusa Heart Study had begun conducting postmortem studies of atherosclerosis in participants in their school-based surveys who died years later from non-CVD causes (Berenson et al. 1992). Risk factor measurements in school, including blood lipids, blood pressure, smoking, and obesity, were associated with the extent and severity of coronary atherosclerosis at death. The findings from these two studies established a critical link between risk factors and pathology at ages well before clinical CVD typically appears. This link is important for strategies of CVD prevention and strengthened evidence for the progressive development of atherosclerosis in early life, as conceptualized decades earlier by Holman and colleagues (1958).

52.3.2.3 Forecasts of the Global CVD Burden

The global dimension of CVD is increasingly recognized, especially in low- and middle-income countries, and misconceptions about the epidemiological transition are beginning to be overcome. For example, the World Heart Federation (WHF) assessed the global situation in a report that addressed the need for international co-operation, the role of the WHF, the profile and magnitude of the global CVD burden, current activities and resources, and strategies for global action (Chockalingam and

Table 52.2 Estimated numbers of deaths and numbers of DALYs lost due to ischemic heart disease and cerebrovascular disease by WHO region, 2001 (Source: World Health Organization 2002)

	WHO region						
Measure of burden	AFR	AMR	EMR	EUR	SEAR	WPR	Total
Deaths (×1,000)							
Ischemic heart disease	333	967	523	2,423	1,972	963	7,181
Cerebrovascular disease	307	454	218	1,480	1,070	1,926	5,455
DALYs lost (×1,000)							
Ischemic heart disease	3,258	6,506	5,353	16,000	20,236	7,373	58,726
Cerebrovascular disease	3,318	4,057	2,364	10,443	9,952	15,736	45,870

Balaguer-Vintró 1999; see also World Heart Federation 2012). It compiled results of surveys from member WHF countries concerning existing policies, programs, and guidelines for CVD prevention, resource requirements and funding sources, human resources, and cardiology infrastructure and practices. Data for 1990 for established market economies, economies in transition (chiefly the countries of the former USSR and others in Eastern Europe), and developing countries demonstrated that the latter contribute far the greatest share of deaths from CVD. This reflects both the virtually ubiquitous occurrence of CVD and the large contribution to the world population from the developing countries.

The World Health Report 2002: Reducing Risks, Promoting Healthy Life provided estimates of both death and disability (in numbers of disability-adjusted life years, or DALYs[1] lost), attributable to IHD and to stroke for the year 2001 (World Health Organization 2002; see also www.who.int; data compiled by the author from Annex Table 2, pp 188–189) for each of the six WHO regions (AFR = Africa, AMR = the Americas, EMR = Eastern Mediterranean, EUR = Europe, SEAR = South-East Asia, and WPR = Western Pacific). Although each region is subclassified by mortality pattern in the source tables, data are summarized here for entire WHO regions, for ischemic heart disease and cerebrovascular disease (Table 52.2).

Of all deaths worldwide, 22.3% were estimated to be due to these two conditions, as were 7.1% of all DALYs lost. Each region contributed substantial numbers to these measures of the CVD burden. Relative contributions of IHD and stroke varied strikingly among regions: Stroke predominated in the Western Pacific, IHD and stroke contributed about equally in Africa, and IHD exceeded stroke in all other regions. The WHO report also provided an extensive analysis of the major risks to health, strategies to reduce these risks, and needs for policies and action to prevent risks.

Through a joint undertaking of WHO, the World Bank, and the Harvard University School of Public Health, the Global Burden of Disease Study first estimated incidence, prevalence, mortality, and disability related to more than 100 diseases and causes of injury as of the year 1990 and projected the global disease burden

[1] 1 DALY lost = 1 full year of life expectancy lost, 2 years lived with 50% disability, etc.

to the year 2020 (Murray and Lopez 1996). IHD and stroke were identified as leading causes of death and disability worldwide in both 1990 and 2020. Among the many findings of importance in the CVD projections were, for change in total CVD death rates from 1990 to 2020: world, +16.2%; established market economies, +1.8%; economies in transition, +19.4%; and developing countries, +28.2%. Again, the relative disadvantage of the developing countries is apparent.

This point was further emphasized in the report, *A Race Against Time: The Challenge of Cardiovascular Disease in Developing Economies* (Leeder et al. 2004). Comparative forecasts of the CVD burden among Brazil, China, India, Russia, and South Africa from 2000 to 2030 demonstrated a significant potential loss by death and disability from the working age population of each country over this period, absent major new efforts in CVD prevention. Loss of workforce translates into slowing of economic and social development, considerations that add strongly to the case for investing in prevention, especially in such countries. Calls for global action from the International Heart Health Society are reflected in their synthesis of successive declarations, *International Action on Cardiovascular Disease: A Platform for Success* (International Heart Health Society 2005).

The worldwide impact of CVDs continues to be assessed, as in the ongoing Global Burden of Disease Project, last reported in 2006 and scheduled for a further update in the near future (Mazzati et al. 2006). Yet the magnitude of the challenge for developing countries is an increasingly dominant theme, well illustrated by World Bank's *Disease Control Priorities in Developing Countries*, in its second edition, and by a report from the Institute of Medicine, *Promoting Cardiovascular Health in the Developing World: A Critical Challenge to Achieve Global Health* (Jamison et al. 2006; IOM 2010a). This concern is intensified in the context not only of CVDs but of NCDs together, as discussed below.

52.4 Determinants

52.4.1 Overview

Many factors contribute to development of CVD. They may be as remote as conditions of fetal development or as immediate as precipitators of plaque rupture in a coronary artery. They may be as distal in their influences as community-level poverty and lack of education or as proximate as an individual's blood pressure. "Determinants," "risk factors," and related terms are often used interchangeably. Here, "determinants" is generally used for the whole array of these influences, from social determinants (such as income or relative poverty, education, and housing quality) to personal risk factors (such as behavioral, metabolic, and physiologic characteristics). Each determinant can potentially impact both CVD rates in populations and CVD risks in individuals. For example, the social determinants (which are population characteristics) affect individuals, and the risk factors (which are individual traits) have their corresponding distributions and trends in populations. The major established risk factors for CVD are highlighted briefly here, with the origins of evidence for these factors discussed elsewhere (Stamler 1992).

A coherent orientation to determinants of CVD begins by distinguishing environmental from genetic factors; addresses variation in CVD rates and risks by demographic characteristics; proceeds by describing the major underlying factors of diet and physical activity, followed by obesity as an outcome of dietary imbalance; considers physiologic consequences of these factors (adverse blood lipid profile, high blood pressure, and diabetes); and concludes with other personal characteristics and social and physical environmental factors. The current epidemiology of each is addressed chapter by chapter in a recent review (Labarthe 2011).

52.4.2 Illustrations

52.4.2.1 Genes and Environment

Methods of genetic, or genomic, epidemiology and its public health implications are presented clearly in recent texts (e.g., Khoury et al. 2004; see also chapter ►Statistical Methods in Genetic Epidemiology of this handbook). Applications to prevention and treatment of CVD were reviewed in a scientific statement of the American Heart Association, with the conclusion that to date the knowledge of greatest utility in this area is that of family history (Arnett et al. 2007). Long-standing awareness of familial resemblance in characteristics related to CVD supports this view, though uncommon conditions such as familial hypercholesterolemia have well-established genetic bases. Meanwhile challenges to study of such "complex" diseases as CVD continue to limit progress in contemporary population-level genomic investigations. Evidence reviewed above demonstrates the strong influence of environmental factors on population differences and secular trends in CVD mortality, patterns of variation that cannot be explained by change in population genetics.

52.4.2.2 Age, Sex, and Race/Ethnicity

Universal demographic characteristics such as age, sex, and race/ethnicity are associated with all aspects of CVD – including both rates and risks of cardiovascular events and the distributions of other determinants. (Note: The term "race" as used here, and in the US public health context, is discussed in Chap. 2, Distributions and Disparities, in Labarthe (2011).) Variation in these measures by age and sex is so ubiquitous that data are rarely presented without age-sex specificity. Measures specific to groups defined by race/ethnicity are also customary, because of both prior knowledge of important group differences and concern to identify and address the disparities in CVD burden that typically accompany disadvantage of some racial or ethnic groups. While these characteristics are often regarded as "unmodifiable," as fixed personal traits, they may serve as markers of particular exposures that, through their modification, can reduce the associated risks.

52.4.2.3 Dietary Imbalance

Dietary patterns to maintain optimum cardiovascular health concern balance, both among components of the diet and between total energy intake and expenditure. The concept of dietary imbalance, then, embraces undesirable patterns of both nutrition and physical activity. The nutritional aspects of CVD, including both atherosclerosis and hypertension, focus on specific nutrients (e.g., saturated, mono- and polyunsaturated, and total fats and specific fatty acids; sodium; fiber; sugar-sweetened beverages; and others) as well as on overall dietary patterns. The American Heart Association, for example, has defined ideal cardiovascular health to include a dietary pattern whose indicators for population monitoring are intakes of fruits and vegetables, fish, whole grains (generally to be increased), sodium, and sugar-sweetened beverages (to be decreased) (Lloyd-Jones et al. 2010). Methods for assessing population-level nutrient intakes are well developed but difficult to implement in large population studies due to constraints of cost and feasibility. The fundamental role of nutrition in CVD and other NCDs warrants substantially increased investment in monitoring population intakes and the policy and environmental contexts that determine food production, distribution, accessibility, price, and choice (see chapter ▶Nutritional Epidemiology of this handbook).

52.4.2.4 Physical Inactivity/Sedentariness

Physical inactivity, the other component of dietary imbalance, contributes to CVD risk through the now-commonplace failure to match energy expenditure with energy intake. Numerous biological mechanisms related to cardiac metabolism and physiology are affected. As a consequence of reduced physical work for personal locomotion and physical exertion in most occupations, research attention turned from occupational to leisure time physical activity (Blackburn 1983). But leisure-time, like occupational activity, has increasingly become devoted to sedentary pursuits (such as watching television), so that average population levels of physical activity have become very low in many countries (President's Council on Physical Fitness and Sports 1996). Standardized instruments have long been available to assess physical activity in population studies, with recent development of the International Physical Activity Questionnaire (IPAC) that permits valid comparisons among countries (Guthold et al. 2008). Physical activity corresponding to "ideal cardiovascular health" equals at least 150 minutes (min) per week of moderate or moderate plus vigorous activity, or 75 min of entirely vigorous activity, for adults; for children the standard is at least 60 min of vigorous activity every day (Lloyd-Jones et al. 2010) (see also chapter ▶Physical Activity Epidemiology of this handbook).

52.4.2.5 Obesity

A direct consequence of the energy component of dietary imbalance is excessive weight gain, measured in its simplest form by the relation of weight to height, expressed as the body mass index (BMI, weight in kilogram/square height in square meter). The World Health Organization classifies BMI levels as chronic energy deficiency (BMI less than 18.5), normal (BMI 18.5–25), or obese (grade 1, BMI 25–30; grade 2, 30–40; grade 3, 40 and greater) (World Health Organization Study Group 1990). Rising prevalence of obesity, beginning in childhood, is a growing global health concern, because of its role in development of multiple NCDs and other determinants of CVD – adverse blood lipid profiles, excessive blood pressure and blood glucose, and type 2 diabetes at ever-earlier ages, now including adolescence (World Health Organization 1999; Koplan et al. 2005; Daniels et al. 2005; International Obesity Task Force 2006) (see chapter ▸Epidemiology of Obesity of this handbook).

52.4.2.6 Adverse Blood Lipid Profile

Early studies in CVD epidemiology identified high blood cholesterol concentration as a "risk factor" for CHD (e.g., Kannel et al. 1961). Over the following decades, deeper insight into lipid metabolism and its relation to atherosclerosis established a more complex role of blood lipids, including high concentrations of low-density lipoprotein (LDL) cholesterol and triglycerides and low concentrations of high-density lipoprotein (HDL) cholesterol. Each of these components of the blood lipid profile was found to predict CVD events and interventions to modify these factors reduced risk. Because total cholesterol concentration is readily tested in non-fasting blood samples, it is a relatively convenient measure for population surveys and is most widely reported. On this basis it has been described as accounting for 45% of IHD mortality and 13% of cerebrovascular (stroke) mortality worldwide (Mazzati et al. 2006).

52.4.2.7 High Blood Pressure

High blood pressure also became established in the earliest epidemiological studies of CVD as a major risk factor for both CHD and stroke. Measured reliably through training of observers in standard auscultatory methods, blood pressure was consistently found to predict CVD events, and treatment of individuals with high blood pressure (and over time, with successively lower starting levels) to reduce the risks. The estimated contributions of high blood pressure to the worldwide burden of IHD and stroke are 45% and 54%, respectively – indicating a particularly serious impact on stroke (Mazzati et al. 2006). High blood pressure results in part from dietary imbalance, through obesity, but is also a specific consequence of high sodium intake. For this reason, national and global health recommendations for CVD prevention place a prominent emphasis on population-wide sodium reduction through changes in food production, preparation, and consumption, as well as changes in health-care systems to improve long-term management (World Health Organization 2007; Asaria et al. 2007; IOM 2010b, c).

52.4.2.8 Diabetes and the Metabolic Syndrome

Diabetes mellitus, especially its predominant form of "type 2 diabetes mellitus," constitutes one of the major NCDs in itself (Narayan et al. 2011). However, it is included here because of its relation to other determinants of CVD and its own contribution to risk of CVD events. "Diabetes" as used here represents a complex set of conditions concerning dietary patterns, energy balance, obesity, production of insulin and its function in glucose metabolism, and interrelations of these factors with blood lipids and blood pressure to comprise what has come to be known as "the metabolic syndrome." Globally, changes in patterns of nutrition and physical activity are regarded as the main drivers of increasing prevalence of diabetes (Yach et al. 2006). Interventions to reduce excess weight through improved diet and physical activity have been shown to prevent progression from the precursor stage of "prediabetes" to diagnosed diabetes and to improve blood lipid and blood pressure levels as well (Diabetes Prevention Program Research Group 2002) (see chapter ▸Epidemiology of Diabetes of this handbook).

52.4.2.9 Smoking and Other Tobacco Use

Cigarette smoking became recognized as a cause of lung cancer in the mid-1960s; only several years later was its causal link with CVD clearly established. More recently still, risk to non-smokers from exposure to tobacco smoke ("secondhand smoke") has become recognized as increasing the risk of acute coronary events (US Department of Health and Human Services 2006). Altogether, the global public health impact of smoking and other tobacco use is in fact several times greater for CVD than for cancer. Smoking becomes a habitual, addictive behavior typically in adolescence, in consequence of the marketing and promotional practices of the tobacco industry. Once addicted, smokers have great difficulty in quitting the habit, and prevention is therefore a high priority in addition to support for smoking cessation (Centers for Disease Control and Prevention 2007). Preventive policies emphasize price increases through taxation to discourage purchase of tobacco products. Systems to facilitate quitting include "quitlines" for telephone contact with personnel trained to support the individual smoker through withdrawal and other obstacles to successful quitting. Community-level or broader laws or regulations promote clean air by restricting smoking in public places, worksites, and other points of exposure (Glantz 2008). Globally, the WHO Framework Convention for Tobacco Control includes many provisions to enable countries to combat the pressures of the tobacco industry to increase the marketing of their products (World Health Organization 2003, 2008). The March 2012 issue of the journal *Tobacco Control* was devoted to a 20-year retrospective on progress in this field and presents numerous perspectives of interest (Tobacco Control 2012).

52.4.2.10 Other Personal Characteristics

Many other personal characteristics have been associated with CVD risk – a 1981 report listed 246 of them, each having been found significantly related to CHD in at least one epidemiological study (Hopkins and Williams 1981). Clearly that number would be far greater today. Three such factors or groups of factors stand out: alcohol consumption, adverse psychological factors, and hemostatic factors. *High alcohol intake* is adversely related to both CHD and stroke. However, the lowest risk of CHD has regularly been found not at the lowest level, but at so-called moderate intake. Interpretation of this pattern has been discussed extensively, and some have advocated low-level alcohol consumption for CVD prevention; official guidelines and recommendations generally avoid such recommendations, rather admonishing those who do use alcohol to do so in moderation, in view of the numerous potential adverse consequences of alcohol use. *Adverse psychosocial factors* of particular interest currently range from feelings of hostility as triggers for acute coronary events to chronic excess of demands over control of one's situation in the workplace. Depression has also become a recognized factor in occurrence of CVD, as has lack of social support from the network in which one lives. *Hemostatic factors* concern the complex interactions of pro- and anti-thrombotic factors that influence formation and resolution of blood clots in the circulation. For example, widespread recommendations for use of aspirin in prevention of CVD events reflect the current understanding of this area. Among other areas of extensive research are inflammation, "novel" markers of risk, and predictive models to assess CVD risk based on multiple baseline factors.

52.4.2.11 Social and Physical Environment

The social and physical environment is distinct from all the foregoing determinants, which are personal characteristics; however, the latter characteristics are largely conditioned by the social and physical environment. The social environment encompasses such aspects as social structure, population diversity, distributions of income and education, and occupational types (World Health Organization 2006). The physical environment may be considered in terms as specific as fine particle air pollution or as broad as the food environment, neighborhood characteristics, transportation systems, walkability as a characteristic of the built environment, and others. These aspects are implicit in Fig. 52.1 above where, in "the present reality," they are reflected in the adverse social and environmental conditions and resulting adverse behavioral patterns that contribute to development of the major CVD risk factors. Alternatively, "a vision of the future" in the figure contemplates social and environmental conditions favorable to health and behavioral patterns that promote health, with consequently reduced risk and rates and severity of CVD events in the population (US Department of Health and Human Services 2003). These determinants of health are addressed in some detail in chapter ►Social Epidemiology of this handbook).

52.5 Causation and Prevention

52.5.1 Overview

Given the plethora of factors possibly associated with the development of CVD, only a few of which are discussed above the question might be posed: Which of these are causal factors? Identifying causes on the basis of associations found in observational epidemiological studies generally rests on the outcome of what might be termed the "differential diagnosis of causation," in three essential steps (Labarthe 2011): (1) Critical assessment identifies apparent or suspected flaws, bias, or confounding that may underlie a spurious association (see relevant chapters in Part II (Methodological Approaches in Epidemiology) of this handbook); (2) statistical assessment addresses the potential role of chance that may produce a chance association (see relevant chapters in Part III (Statistical Methods in Epidemiology) of this handbook); and (3) consideration of several properties of the evidence supports judgment whether an association is best interpreted as causal. The properties to be considered were elaborated through a series of commentaries appearing in the *Journal of Chronic Diseases* in the late 1950s, culminating in a contribution by Sartwell (1959). They became codified especially in the 1964 report on smoking and health in the USA and the paper by Hill that followed (Advisory Committee to the Surgeon General of the Public Health Service 1964; Hill 1965). To the Advisory Committee list, traceable to Sartwell and predecessors (consistency, strength, specificity, temporal relation, and coherence), Hill's list added four others (biological gradient, plausibility, experiment, and analogy).

Such differential diagnosis of causation is not always applied, however, and how many of the 246 factors listed by Hopkins and Williams, cited above, might have passed this more stringent procedure, then or now, is unknown. Even the first reported use of the term "risk factor" noted that evidence was insufficient to conclude that the associations described were causal and that further study was needed (Kannel et al. 1961). Thus, it would be inappropriate to interpret references to "risk factors" throughout the literature of CVD epidemiology as necessarily meaning "causal factors" in the strict sense suggested above. Further, what constitutes a causal explanation of associations between exposures and diseases depends in part on the question being asked. Different questions imply different concepts of causation and require different kinds of causal answers, as follows.

52.5.2 Illustrations

52.5.2.1 Concepts of Causation

One Cause or Many? Conditions such as CHD and stroke are generally accepted to be consequences of multiple factors acting in concert ("multifactorial causation"). Admitting multiple risk factors, however, may deny credibility to any one of them (McCormick and Skrabanek 1988; Stehbens 1992). This view stems from the ancient concept of one "necessary and sufficient cause" as the whole explanation of a phenomenon. This concept was generally abandoned for the chronic diseases during the 1950s and 1960s in the effort to explain the consistently incomplete conjunction of any one exposure and disease. That is, exposure and disease overlap when they are causally related, but neither does every case result from the same exposure nor does exposure uniformly produce the same disease. Models of CVD causation have advanced by taking into account the multivariate setting, in which single causal factors operating together tend to multiply risk above that attributable to the sum of effects of the individual factors. As a result, multivariate risk scores have been calculated for use both as the outcome variable in intervention studies and as a tool for assessing risk of a future CVD event as a basis for deciding whether treatment is warranted (Farquhar et al. 1990; De Backer et al. 2003). Currently the multifactorial view strongly prevails.

One Outcome or Many? A theme of CVD epidemiology in its early years, rediscovered only recently, was that increased susceptibility resulting from any social factor should not be expected to be narrowly specific in its disease expression. Therefore, factors increasing the risk of CHD should be studied as factors in illness generally and not for CHD alone. In this view, causes could have quite varied outcomes, so that a causal model specific to one disease outcome must be inherently incomplete and therefore subject to misinterpretation or misunderstanding. Echoing this theme, Stallones (1980) argued that models of causation that converge on a single disease outcome should be superseded by considering "the interdependence of a number of diseases, characteristics of individuals, and environmental and social variables as elements in a constellation which is *n*-dimensional" (p 76). The tendency to focus on single diseases is yielding in some degree to embracing the NCDs, in part because of the recognized contribution of causal factors for CVDs to other NCDs as well. Wider appreciation of the "*n*-dimensional" constellation of elements in a causal framework is also consonant with the recent resurgence of concepts of "social determinants of health" and "population health."

Immediate Causes or "Fundamental Causes?" Here the argument is that a focus on immediate or proximal causes of disease, and interventions to mitigate their effects, is inherently flawed. For example, high blood pressure may be causally associated with rates and risks of stroke, but to concentrate public health efforts on detection and control of high blood pressure is to miss the greater public health need. This need is to address population characteristics such as poverty, illiteracy, unemployment, and other social ills that would diminish health and well-being by countless other mechanisms, even if high blood pressure was completely removed as a public health problem. The differential diagnosis of causation includes specificity of an association as one consideration, in the sense that an association which uniquely links a particular exposure with a particular disease may more reasonably be thought causal than one in which both the exposure and the disease are related to multiple other conditions. The "fundamental causes" are in principle non-specific and partly for that reason fail to satisfy narrower, more immediate, and disease-specific concepts of "cause."

All Causes or Only a Few? The all-inclusive view of CVD causes recognizes that innumerable factors influence risk, some in adverse and others in protective ways. This view finds excitement in the "emerging risk factors" whose novelty and interest reinforce confidence in the advances of science and the sense of being at the leading edge. In this view, CVD epidemiology may be at the threshold of an era of unprecedented discovery with the advent of population genomics. One causal model for CHD portrays, in one tier, simply "genes"; an indefinitely large array of individual genes is linked through multiple connections from the first to the second tier, a "network of intermediate traits" (hemostasis, lipid metabolism, carbohydrate metabolism, and blood pressure regulation); and these four traits converge to act on the third tier, the "probability of CHD," which is represented graphically as a surface that represents age and environment (Sing et al. 1992). The biological complexity depicted by this scheme represents a rich body of knowledge whose growth is almost inevitable as science advances at all three levels and the connections between them are continually elaborated. And this represents only the convergence of factors on one outcome, not the "*n*-dimensional array" conceived by Stallones (1980).

The "only a few" point of view focuses on only those factors regarded as useful for widespread intervention to arrest and reverse the CVD epidemic around the world. Those few factors, variously termed the major or established or traditional risk factors, consistently include blood lipids, blood pressure, and smoking. Stamler 1992 identifies four factors: "... 'rich' diet, diet-related above-optimal levels of serum total cholesterol (TC) and blood pressure (BP), and cigarette smoking." He explains (p 36):

> All four of these risk factors are designated *established* because substantial amounts of data from many disciplines have demonstrated their significant role in the etiology of epidemic CHD. They are designated *major* for three reasons: their high prevalence in populations, particularly in Western industrialized countries, their strong impact on coronary risk, and their preventability and reversibility, primarily by safe improvements in population lifestyles, from early childhood on. (Age and male gender are known risk factors, but are not amenable to influence and hence are not designated major; diabetes is a known risk factor, but of lower prevalence than cigarette smoking and elevated TC and BP in most populations, and hence is not designated major.) ... "Rich" diet is pivotal among these four – the primary and essential cause of the coronary epidemic.

These two views, all versus few, are commonly seen as conflicting. From the "all" point of view, the "few" perspective ignores evidence for many new and emerging factors and fails to see the potential of genetic epidemiology, for example, to provide invaluable and perhaps radically new insights about causation and prevention of CVD. On the other hand, from the "few" perspective, preoccupation with these factors is a significant and unwelcome distraction from the needed focus on applying knowledge of the major risk factors and research to make that application as effective as possible for prevention. Perhaps from the perspective of "fundamental causes," this polarity of views is irrelevant, as the main issues are the broadest social determinants of health and quality of life, of which CVD is only one aspect.

A resolution of these seemingly conflicting views is found in the point made by Stallones (1980), who suggested that models of causation depend for their value on their utility and that utility depends in turn on the purpose for which a model is used. If the dual purposes of intellectual pursuit and public health practice are accepted as complementary in CVD epidemiology and prevention, the views of "all" and "few" can both be embraced. "Epidemiology and public health have the luxury as well as the necessity of entertaining both views of causation in pursuit of the complementary purposes of advancing knowledge and serving the health of the public" (Labarthe 1998, p 463).

Causes of Cases or Causes of Incidence? This last contrast addresses a distinction between two kinds of causal questions posed by Rose (1985, p 33):

> The first seeks the causes of cases, and the second seeks the causes of incidence. "Why do some individuals have hypertension?" is a quite different question from "Why do some populations have much hypertension, whilst in others it is rare?" The questions require different kinds of study, and they have different answers.

The distinction emphasizes that what appear to be causal factors (from studies of individual risks within populations) may instead be only markers of susceptibility to true causal factors to which exposure is very common throughout the population and that detection of true causal factors requires comparison of populations with different incidence rates to identify the determinants of those differences. As with each of the preceding contrasts, this distinction relates closely to concepts of prevention and has direct implications for thinking about individual and population-wide levels of intervention.

52.5.2.2 Concepts of Prevention

As the foundation of epidemiological knowledge about CVD expanded, recommended approaches to CVD prevention evolved correspondingly. For example, under the aegis of the former WHO Cardiovascular Disease Unit, expert meetings were convened on many occasions from the 1960s through the mid-1990s to address CHD prevention, primary prevention of hypertension, prevention in childhood and youth of adult CVD, and other topics. Official recommendations and intervention guidelines have been developed by such national organizations as the American Heart Association and American College of Cardiology and the US National Heart, Lung, and Blood Institute and at a regional level by such joint efforts as those of the European Society of Cardiology, the European Atherosclerosis Society, the European Society of Hypertension, and the South Asian Association for Regional Cooperation.

Concepts of prevention reflected in these many recommendations and guidelines often embrace two broad approaches articulated in the early 1980s by Rose (1981) – the individual (or high-risk) approach and the population-wide approach. The first approach addresses reduction of risk for cardiovascular events among persons with already manifest CVD or risk factors, and the second addresses risk factor reduction across the whole population. These two approaches are complementary. "The 'high-risk' strategy of prevention is an interim expedient, needed in order to protect susceptible individuals, but only for so long as the underlying causes of incidence remain unknown or uncontrollable; if causes can be removed, susceptibility ceases to matter" (Rose 1981, p 38).

The high-risk approach has been considerably refined so as to identify, on the basis of multiple factors (such as sex, current smoking, age, systolic blood pressure, and total cholesterol concentration), persons whose combined risk characteristics predict a specified probability of a major cardiovascular event within the ensuing 10 years (De Backer et al. 2003). Based on such scoring, policy decisions are made regarding which interventions will be recommended, and perhaps financed, for which risk levels.

The population-wide approach has evolved to include not only interventions to shift unfavorable distributions of the major risk factors across the whole population but also prevention of adverse levels and distributions of these risk factors in the first place, in populations where this remains possible. This latter concept of "primordial prevention" contemplates prevention of risk factor epidemics themselves as an essential approach to CVD prevention, especially in developing countries (Strasser 1978).

The high-risk approach to risk reduction as conceived by Rose (1981) would today encompass both secondary prevention (applied to individuals already manifesting CHD or stroke) and primary prevention (applied to individuals at high risk but without manifest CVD). The population-wide approach to risk reduction would consist of that part of primary prevention applied to the population as a whole; it may be complemented by prevention of risk factors in the first place (primordial prevention). Broadly these may be contrasted as "remedial" and "primordial" strategies, for CVD prevention and cardiovascular health promotion, respectively (Labarthe and Dunbar 2012).

In relation to the intervention approaches represented in Fig. 52.1, the remedial strategies extend from risk factor detection and control through end-of-life care at the individual level and include the population-level approaches, from policy and environmental change through behavior change, insofar as these are undertaken to reverse already increased risk. The primordial strategies, for health promotion, are also population-wide, undertaken with the intent to preserve cardiovascular health by fostering favorable social and environmental conditions and health behaviors.

The scope of CVD prevention and relevance of this framework for public health action take added meaning from the heightened appreciation of the collective global burden of NCDs, beyond CVDs alone (United Nations General Assembly 2011). Both the case for prevention and estimates of the health impact of strategic interventions gain substantially from this more inclusive perspective.

52.5.3 The Role of Trials

Epidemiological evidence for causation of CVD has come primarily from observational studies and has sometimes been supported indirectly by experimental studies in which exposure to suspected factors has been reduced by intervention. The balance of evidence for causation is heavily weighted toward observational studies. How is evidence for CVD prevention obtained? Evidence from observational studies may be considered sufficient to establish causation, but different requirements are often, perhaps typically, applied to prevention – primarily experimental evidence of the efficacy and safety of a proposed intervention.

This question thus calls attention to the role of trials or, more broadly, intervention studies. These studies encompass a wide variety of study questions and designs and include clinical trials of secondary or primary prevention, population-based trials of primary prevention or health promotion, and community demonstrations of intervention that may include combinations of health promotion, primary prevention, and secondary prevention. It is beyond the scope of this review to detail the evidence for prevention; such commentaries are available elsewhere (Blackburn 1992; Labarthe 2011). It is useful, however, to summarize briefly the general types of intervention studies on which CVD prevention is based (see chapters ▶Intervention Trials, ▶Confounding and Interaction, and ▶Epidemiological Field Work in Population-Based Studies of this handbook):

- The great preponderance of experimental studies in CVD prevention have been *clinical trials in secondary prevention*. These are designed to test the efficacy and safety of individual-level interventions to prevent recurrent CVD events or related outcomes among survivors of first events or persons otherwise known to have preexisting CVD. The typical multicenter organization of such trials, the recruitment and enrollment of hundreds or thousands of participants, and the multi-year periods of follow-up combine to make these studies epidemiological in scale, even though they may be altogether clinically based and not referable to any definable population beyond the participating institutions.
- *Primary prevention trials* are designed to test the efficacy and safety of individual-level interventions, but the participants are initially free of known CVD. The outcome is usually the first qualifying CVD event, be it an acute coronary event, stroke, or other defined outcomes, depending on the question under study. For example, in studies of primary prevention of high blood pressure, progression of blood pressure levels may be the outcome of interest. Study participants may be identified in clinical settings or drawn from defined communities, as volunteers or through systematic sampling, such as from work sites or geographical residential areas. Eligible participants must be apparently well, so only interventions relatively free of known adverse effects or other impediments to long-term adherence to study treatment guidelines are good candidates for evaluation. Because participants' average risk of CVD outcomes is relatively lower in primary than in secondary prevention trials, thousands of participants and follow-up periods of several years are typically needed to observe a sufficient number of outcomes for satisfactory analysis.
- *Community trials* in CVD prevention are used to study intervention in whole populations defined by geographical area of residence, place of work, or otherwise. If insufficient numbers of study units such as schools or factories are involved to permit true random allocation, as is usually the case, the term "quasi-experimental study" is sometimes used. Systematic comparison among similar communities with and without intervention is still implied, even if the number of intervention and comparison units is small. Fatal or non-fatal CVD events could be the primary endpoint in such trials, but these studies are often designed to address risk factors or behavioral outcomes instead. The choice of interventions shares the same considerations as described for primary prevention trials.
- A *community demonstration* of intervention represents a distinct shade of meaning from "community trials" in that intervention may be implemented in a single community, with or without a control community for comparison. Community

demonstrations may combine multiple intervention strategies into a secondary and primary prevention program to influence individuals' practices as well as community-wide aspects whose target is the population as a whole. Evaluating the impact of such an intervention may be based on CVD events, but practical constraints often arise. Evaluation may therefore be limited to the procedures by which the intervention was delivered and such proximate outcomes as changes in behavior or risk factor levels.

Analogous to the question of causation in epidemiological studies is the question whether a given intervention actually produces the intended outcome. The most convincing evidence comes from well-designed and properly conducted randomized trials; however, population-wide public health interventions may only rarely be amenable to evaluation through studies of this kind. This circumstance leads to the seeming dilemma of holding population-wide interventions to a standard of evidence that can rarely, if ever, be attained in practice and therefore giving rise to the perception that population-wide interventions are only rarely, if ever, scientifically justified. However, the essence of the question for a proposed public health intervention is whether it is better supported by all the relevant scientific evidence than is the status quo, as represented by current conditions or policies. If so, intervention would be appropriate if other due considerations were met, and rigorous evaluation of the intervention's effect once implemented could provide further evidence as to its benefit, cost-effectiveness, and long-term feasibility.

Among a great many potential examples of intervention studies, one with broad implications for prevention policy is especially convenient for illustration. The North Karelia Project provides a classic example; detailed published accounts of its organization, design, implementation, and evaluation are provided elsewhere (Vartiainen et al. 1994; Puska et al. 1998; Puska 2002).

As shown in Fig. 52.2, Finnish men experienced exceptionally high CHD mortality that had increased sharply in the 1950s. Since 1959, studies in Eastern and Western Finland as part of the Seven Countries Study had documented the problem in some detail and confirmed the highest CHD rates to be in Eastern Finland. Concern about this led to implementation in 1972 of a multifaceted community-based prevention project in which North Karelia (population 210,000) and Kuopio (population 250,000), both in Eastern Finland, would be intervention and control communities, respectively. Among the many components of the project were programs targeting high blood cholesterol concentration, high blood pressure, and smoking. Extensive community involvement and engagement with health services were major aspects of these programs. Risk factor distributions and other characteristics in both communities were assessed by sample surveys among adults aged 30 to 59 years in 1972 and every 5 years until 1992, and mortality was monitored from 1969 to 1992.

Twenty-year changes in risk factors for women included reductions in cholesterol concentration by 18% and in diastolic blood pressure by 13%, while smoking increased from 11% to 25%. Corresponding changes for men were reductions of 13% and 9% and a decrease in smoking from 53% to 37%. Changes in dietary habits and physical activity underlying the change in risk factor distributions have also been described.

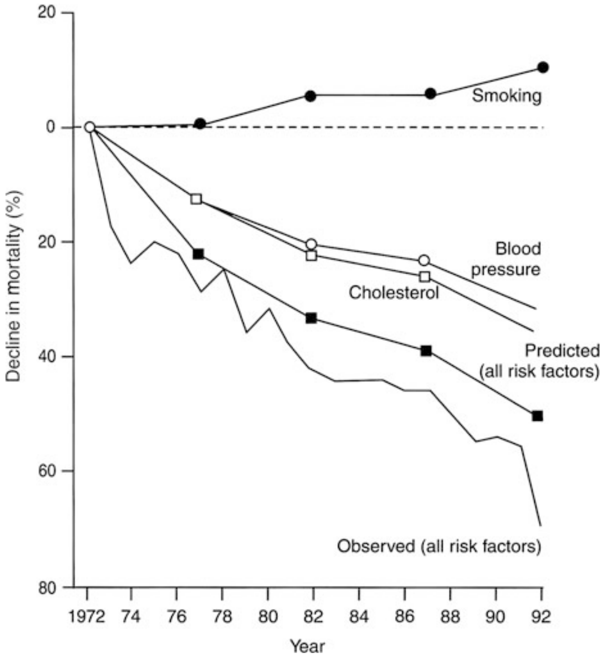


Fig. 52.3 Observed and predicted decline in mortality from ischemic heart disease in women aged 35–64 in Finland (Source: Vartiainen et al. 1994)

Changes in CHD mortality were predicted from the actual risk factor distributions observed in the serial cross-sectional surveys. These results are displayed in Fig. 52.3 for women and Fig. 52.4 for men. The predicted mortality effects of change in each risk factor separately and together are shown in each figure, as well as the observed decrease in mortality (by 68% for women and 55% for men) which closely paralleled but actually exceeded the cumulative predicted changes.

What was the contribution of intervention to these changes? Changes in risk factors were greater in the intervention area, North Karelia, than in the control area, Kuopio, only in the first 5 years; thereafter, they were similar. Change in mortality was not limited to the study areas: Overall, in the whole of Finland, CHD mortality declined by 50% from 1970 to 1992 and stroke mortality declined markedly. Interpretation of the effect of intervention in North Karelia is conditioned by the study design (comparison of two selected communities and not a randomized trial in a large number of communities), but the investigators addressed these issues and supported the conclusion that the mortality decline was a direct consequence of the risk factor changes (Vartianinen et al. 1994; Puska et al. 1998). That the project began to influence policy nationwide after its first 5 years probably accounts for the broader evidence of risk factor and mortality change in Finland as a whole that has continued to the present and makes it difficult to isolate the intervention effects to North Karelia alone. It was recognized that improvements in treatment of coronary

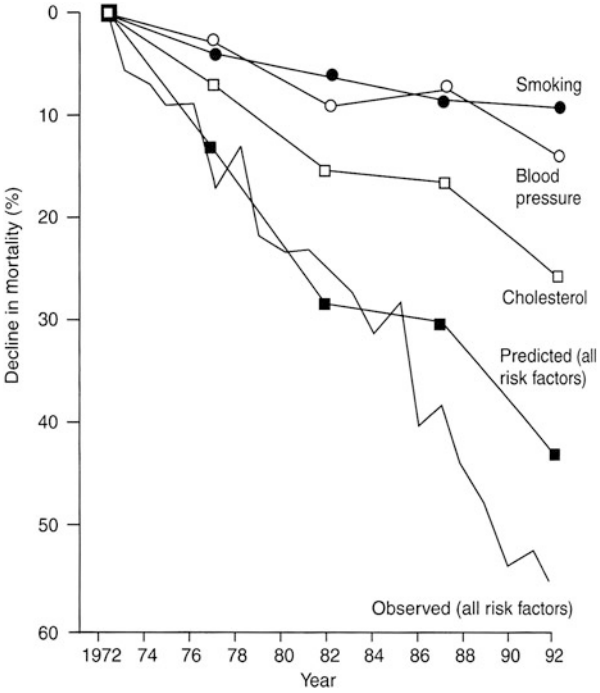


Fig. 52.4 Observed and predicted decline in mortality from ischemic heart disease in men aged 35–64 in Finland (Source: Vartiainen et al. 1994)

events could also have contributed, a factor that would be evaluated further with Finland's participation in the WHO MONICA Project.

From the perspective advanced by Rose (1981), the Finnish experience supports the view that the major determinants of incidence have been identified, and population-wide efforts to modify these determinants have made it possible to prevent the mass occurrence of CVD in whole populations. At the broadest level of community change and population-wide intervention, coupled with health-care system interventions to improve services for risk factor detection and control, the North Karelia Project is a powerful demonstration of the potential for an integrated, coordinated, and sustained public health effort to affect the major cardiovascular conditions of our time, CHD and stroke.

52.6 Research Needs

This chapter has focused on knowledge of the epidemiology of CVD based on decades of research through application of the methods outlined above, under "Scope and Basic Concepts." It would be incomplete without reference to current and future research needs, in terms of potential topics as well as infrastructure requirements. Among the many topics that could be considered in relation to research priorities in CVD prevention and cardiovascular health promotion, from the public health perspective, two questions are selected for emphasis here, with additional attention to questions of research capacity and organizational arrangements for research.

52.6.1 How Can Healthy Dietary Patterns for All Be Achieved?

Food production, distribution, accessibility, price, and choice must be aligned with nutritional requirements for good health, including cardiovascular health, if the burden of CVD and other NCDs is to be reduced, from local to national, regional, and global levels. Research is needed to monitor changes in the food supply, content of specific food products (e.g., sodium), and patterns of food consumption at the population level, over time. The roles and strategies of multinational food corporations must be better understood and anticipated in order for effective policy initiatives to be developed by responsible governmental agencies and interested private organizations. At the same time, factors that more immediately influence what people consume – within the range of available choice – must also be considered. In sum, research is needed to evaluate comprehensively policies, programs, and practices designed to overcome widely prevalent dietary imbalance. To the extent that this is unfamiliar ground to epidemiologists, collaboration with others is required.

52.6.2 How Can Cardiovascular Health Be Preserved from Childhood Onward?

Atherosclerosis begins in childhood and youth, and the major risk factors – high cholesterol concentration, high blood pressure, and smoking – contribute beginning from adolescence or earlier (Berenson 1986; Mahoney et al. 1991). Further, retaining low CVD risk at middle age greatly reduces morbidity in later years (Stamler et al. 1999; Lloyd-Jones et al. 2006). Yet this knowledge has failed to stimulate the needed programs on a sufficient scale to avert development of risk factors and subclinical disease in the whole population. Guidelines to address CVD risk factors in early life have been published repeatedly, but few large-scale programs have been mounted to assess the impact of the recommended interventions (World Health Organization Expert Committee 1990). The recent focus of the American Heart Association and others on improving cardiovascular health of the population has called attention to the measurable *loss of cardiovascular health* beginning in childhood (Roger et al. 2011). Therefore, the need to continue to advance policy development in CVD prevention seems clear, and research in this area – beginning with implementation and evaluation of currently available approaches – is a high priority.

New approaches, building on recent work in positive health, may offer additional insights into potential interventions to preserve cardiovascular health from childhood onward (Boehm and Kubzansky 2012). The question of how important maternal and fetal influences are on development of CVD, from risk factor levels in childhood to overt CVD in adults, continues to be debated (see Barker 1992, 1998). In view of the potential impact on health generally as well as CVD in particular, resolving this question by conducting trials of optimum nutritional and other supportive maternal care, coupled with postnatal nutrition to evaluate benefits on metabolic and physiologic mechanisms involved in later risk factor development, is important.

52.6.3 How Can Research Capacity Be Increased?

The need for applied research in CVD epidemiology far exceeds current capacity. This is due in part to the persistence of CVD as leading cause of death throughout the world and forecasts for its continuation in this position at least through the next two decades. Training and engagement in substantial epidemiological research projects and programs, as well as ongoing and expanding activities in monitoring and surveillance, are essential for building this capacity. A model program for such training is illustrated by the International Ten-Day Teaching Seminars on Epidemiology and Prevention of Cardiovascular Diseases, a program now in its 37th year, and sponsored principally by a local host in various countries, year by year (Labarthe et al. 1998). Other international training programs (such as that conducted in developing countries under the auspices of the Department of Social and Preventive Medicine, University of Lausanne) offer additional models, all of which should be replicated many times over to reach much-increased numbers of health professionals. It is of course also necessary to develop funded research to provide settings where trained epidemiologists and others in CVD prevention can do the needed work, extend their skills, and provide mentorship and training to others.

52.6.4 What Organizational Arrangements Are Required?

The role of supporting organizations and agencies in CVD epidemiology and prevention has been critical in the past and may be even more so in the immediate future. WHO, WHF, the World Bank, the Global Forum for Health Research, and other agencies with global reach are essential partners for the kind of network cited by WHO Director-General Brundtland in her proposed global strategy for NCD prevention and control (World Health Organization 1999). National organizations in many countries are also important. In the United States, for example, the National Center for Chronic Disease Prevention and Health Promotion of the Centers for Disease Control and Prevention, Department of Health and Human Services, is committed to global health efforts. This interest is an integral part of *A Public Health Action Plan to Prevent Heart Disease and Stroke* (US Department of Health and Human Services 2003), source of Fig. 52.1 above, which provides for regional and global partnerships in cardiovascular health in pursuit of common aims and in the expectation of mutual gain from sharing information and experience. Out of this *Action Plan*, a public-private partnership organization emerged, the National Forum for Heart Disease and Stroke Prevention (2012, see www.nationalforum.org). Guided by this framework, the National Forum sets priorities and advocates for policy development in such areas as addressed here. Potentially through implementation of a nascent Policy Depot for sharing CVD-related policy interests globally, the National Forum can contribute to support of epidemiological research on the global level (Mason et al. 2012).

52.7 Conclusions

Having discussed potential determinants, approaches for prevention, and the priorities in the research of CVD, one question remains to be addressed: *How can epidemiology contribute to societal change?* CVD epidemiology over the past half century has established understanding of the causes of epidemic CHD and stroke as well as strategies to prevent these diseases. The challenge that lies immediately ahead is to apply this knowledge effectively to the benefit of the world's population. Societal change is needed to achieve the goals of longer, healthier life free of significant disability due to CVD and other NCDs and free of disparities in the occurrence of these conditions within and among countries. Societal change will occur regardless whether epidemiology and public health are considered, as change is a continuing – albeit sometimes fitful and unpredictable – process with its own powerful impetus from many sources. But change that is best will be informed and influenced by knowledge of consequences for CVD, for other NCDs, and for health generally. Positive change can be accomplished best to the extent that epidemiologists and others who hold that knowledge contribute to the direction of this change.

References

Advisory Committee to the Surgeon General of the Public Health Service (1964) Smoking and health: report of the Advisory Committee to the Surgeon General. Public Health Service, US Department of Health, Education and Welfare, Atlanta, GA

Arnett DK, Baird AE, Barkley RA, Basson CT, Boerwinkle E, Ganesh SK, Herrington DM, Hong Y, Jaquish C, McDermott DA, O'Donnell CJ (2007) Relevance of genetics and genomics for prevention and treatment of cardiovascular disease: a scientific statement from the American Heart Association Council on Epidemiology and Prevention, the Stroke Council, and the Functional Genomics and Translational Biology Interdisciplinary Working Group. Circulation 115:2878–2901

Asaria P, Chisholm D, Mathers C, Ezzati M, Beaglehole R (2007) Chronic disease prevention: health effects and financial costs of strategies to reduce salt intake and control tobacco abuse. Lancet 370:2044–2053

Association of State and Territorial Directors of Health Promotion and Public Health Education (2001) Policy and environmental change: new directions for public health (executive summary). US Centers for Disease Control and Prevention, Atlanta, GA

Barker DJP (1998) Mothers, babies and health in later life, 2nd edn. Churchill Livingston, Edinburgh

Barker DJP (ed) (1992) Fetal and infant origins of adult disease. BMJ Books, London

Berenson GS (1986) Causation of cardiovascular risk factors in children: perspectives on cardiovascular risk in early life. Raven, New York

Berenson GS, Wattigney WA, Tracy RE, Newman WP III, Srinivasan SR, Webber LS, Dalferes ER Jr, Strong JP (1992) Atherosclerosis of the aorta and coronary arteries and cardiovascular risk factors in persons ages 6 to 30 years and studied at necropsy (the Bogalusa Heart Study). Am J Cardiol 70:851–858

Blackburn H (1983) Physical activity and coronary heart disease: a brief update and population view (Part I). J Card Rehabil 3:101–111

Blackburn H (1992) Community programs in coronary heart disease prevention and health promotion: changing community behaviour. In: Marmot M, Elliott P (eds) Coronary heart disease epidemiology: from aetiology to public health. Oxford University Press, Oxford, pp 495–514

Blackburn H (2012) Preventing heart attack and stroke. A history of cardiovascular epidemiology. http://www.epi.umn.edu/cvdepi/. Accessed on 18 July 2012

Blackburn H, Keys A, Simonson E, Rautaharju P, Punsar S (1960) The electrocardiogram in population studies: a classification system. Circulation 21:1160–1175

Bloom DE, Cafiero ET, Jané-Llopis E, Abrahams-Gessel S, Bloom LR, Fathima S, Feigl AB, Gaziano T, Mowafi M, Pandya A, Prettner K, Rosenberg L, Seligman B, Stein A, Weinstein C (2011) The global economic burden of non-communicable diseases. World Economic Forum, Geneva

Boehm JK, Kubzansky LD (2012) The heart's content: the association between positive psychological well-being and cardiovascular health. Psych Bull 138:655–691

Centers for Disease Control and Prevention (2007) Best practices for comprehensive tobacco control programs – 2007. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta, GA

Chockalingam A, Balaguer-Vintró I (eds) (1999) Impending global pandemic of cardiovascular diseases: challenges and opportunities for the prevention and control of cardiovascular diseases in developing countries and economies in transition. Prous Science, Barcelona

Daniels SR, Arnett DK, Eckel RH, Gidding SS, Hayman LL, Kumanyika S, Robinson TN, Scott BJ, St. Jeor S, Williams CL (2005) Overweight in children and adolescents. Pathophysiology, consequences, prevention, and treatment. Circulation 111:1999–2012

Dawber TR, Meadors GF, Moore FE Jr (1951) Epidemiological approaches to heart disease: the Framingham study. Am J Public Health 41:279–286

De Backer G, Ambrosioni E, Borch-Johnsen K, Brotons C, Cifkova R, Dallongeville J, Ebrahim S, Faergeman O, Graham I, Mancia G, Cats VM, Orth-Gomér K, Perk J, Pyörälä K, Rodicio JL, Sans S, Sansoy V, Sechtem U, Siolber S, Thomsen T, Wood D (2003) Executive summary: European guidelines on cardiovascular disease prevention in clinical practice. Third Joint Task Force of European and Other Societies on Cardiovascular Disease Prevention in Clinical Practice. Eur Heart J 24:1601–1610

Diabetes Prevention Program Research Group (2002) Reduction in the incidence of T2DM with lifestyle intervention or metformin. N Engl J Med 346:393–403

Epstein FH (2005) Contribution of epidemiology to understanding coronary heart disease. In: Marmot M, Elliott P (eds) Coronary heart disease epidemiology. From aetiology to public health. Oxford University Press, Oxford, pp 8–17

Farquhar JW, Fortmann SP, Flora JA, Taylor B, Haskell WL, Williams PT, Maccoby N, Wood PD (1990) Effects of community wide education on cardiovascular disease risk factors: the Stanford Five-City Project. JAMA 264:359–365

Freis ED (1990) Reminiscences of the veterans administration trial of the treatment of hypertension. Hypertension 16:472–475

Glantz SA (2008) Meta-analysis of the effects of smokefree laws on acute myocardial infarction: an update (Letter). Prev Med 47:452–453

Goff DG, Howard G, Russell GB, Labarthe DR (2001) Birth cohort evidence of population influences on blood pressure in the United States, 1887–1994. Ann Epidemiol 11:271–279

Gordon T (1957) Mortality experience among the Japanese in the United States, Hawaii, and Japan. Public Health Rep 72:543–553

Guthold R, Ono T, Strong KL, Chatterji S, Morabia A (2008) Worldwide variability in physical inactivity. A 51-country survey. Am J Prev Med 34:486–494

Hill AB (1965) The environment and disease: association or causation? Proc Soc Med 58:295–300

Holman RL, McGill HC, Strong JP, Geer JC (1958) The natural history of atherosclerosis: the early aortic lesions as seen in New Orleans in the middle of the 20th century. Am J Pathol 34:209–235

Hopkins PN, Williams RR (1981) A survey of 246 suggested coronary risk factors. Atherosclerosis 40:1–52

International Heart Health Society (2005) International action on cardiovascular disease: a platform for success based on international cardiovascular disease (CVD). Declarations. International Heart Health Society, Vancouver, Canada

International Obesity Task Force (2006) The Sydney Principles: guiding principles for achieving a substantial level of protection for children against the commercial promotion of foods and beverages. http://www.iotf.org/sydneyprinciples/index.asp. Accessed on 28 June 2009

INTERSALT Co-operative Research Group (1986) INTERSALT study: an international co-operative study on the relation of blood pressure to electrolyte excretion in populations. I: Design and methods. J Hypertens 4:781–787

IOM (Institute of Medicine) (2010a) Promoting cardiovascular health in the developing world: A critical challenge to achieve global health. The National Academies Press, Washington, DC

IOM (Institute of Medicine) (2010b) A population-based policy and systems change approach to prevent and control hypertension. The National Academies Press, Washington, DC

IOM (Institute of Medicine) (2010c) Strategies to reduce sodium intake in the United States. The National Academies Press, Washington, DC

Jamison DT, Breman JG, Measham AR, Alleyne G, Claeson M, Evans DB, Jha P, Mills A, Musgrove P (eds) (2006) Disease control priorities in developing countries, 2nd edn. International Bank for Reconstruction and Development/The World Bank, Washington, DC

Kagan A, Yano K (1996) Ni-Hon-San Study. In: Kagan A (ed) The Honolulu heart program: an epidemiological study of coronary heart disease and stroke. Harwood Academic, Amsterdam, pp 21–32

Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J III (1961) Factors of risk in the development of coronary heart disease – six-year follow-up experience: the Framingham Study. Ann Intern Med 55:33–50

Keys A (1980) Seven Countries: a multivariate analysis of death and coronary heart disease. Harvard University Press, Cambridge, MA

Khoury MJ, Little J, Burke W (eds) (2004) Human genome epidemiology: a scientific foundation for using genetic information to improve health and prevent disease. Oxford University Press, Oxford

Koplan JP, Liverman CT, Kraak VI (eds) (2005) Preventing childhood obesity: Health in the balance. Committee on Prevention of Obesity in Children and Youth, Food and Nutrition Board, Board on Health Promotion and Disease Prevention, Institute of Medicine. The National Academies Press, Washington, DC

Labarthe DR (1998) Epidemiology and prevention of cardiovascular diseases: a global challenge. Aspen, Gaithersburg, MD

Labarthe DR (2011) Epidemiology and prevention of cardiovascular diseases: a global challenge, 2nd edn. Jones and Bartlett, Sudbury MA

Labarthe DR (2013) Cardiovascular health and disease. Oxford bibliographies in public health. McQueen, David V, Ed. New York: Oxford University Press. www.oxfordbibliographies.com

Labarthe DR, Dunbar SB (2012) Global cardiovascular health promotion and disease prevention: 2011 and beyond. Circulation 125:2667–2676

Labarthe DR, Khaw K-T, Thelle D, Poulter N (1998) The Ten Day International Teaching Seminars on Cardiovascular Epidemiology and Prevention: a 30-year perspective. CVD Prev 1:156–163

Leeder S, Raymond S, Greenberg H (2004) A race against time: the challenge of cardiovascular disease in developing countries. The Trustees of Columbia University in the City of New York, New York

Lloyd-Jones DM, Leip EP, Larson MG, D'Agostino RB, Beiser A, Wilson PWF, Wolf PA, Levy D (2006) Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. Circulation 113:791–798

Lloyd-Jones DM, Hong Y, Labarthe D, Mozaffarian D, Appel LJ, Van Horn L, Greenlund K, Daniels S, Nichol G, Tomaselli GF, Arnett DK, Fonarow GC, Ho PM, Lauer MS, Masoudi FA, Robertson RM, Roger V, Schwamm LH, Sorlie P, Yancy CW, Rosamond WD, and on behalf of the American Heart Association Strategic Planning Task Force and Statistics Committee (2010) Defining and setting national goals for cardiovascular health promotion and disease reduction. The American Heart Association's strategic impact goal through 2020 and beyond. Circulation 121:586–613

Luepker RV, Evans A, McKeigue P, Reddy KS (2004) Cardiovascular survey methods. World Health Organization, Geneva

Mahoney LT, Lauer RM, Lee J, Clarke WR (1991) Factors affecting tracking of coronary heart disease risk factors in children. The Muscatine Study. Ann NY Acad Sci 623:120–132

Marmot M, Elliott P (eds) (2005) Coronary heart disease epidemiology: from aetiology to public health, 2nd edn. Oxford University Press, Oxford

Mason K, Chockalingam A, Prudhomme S, Stachenko S, Pearson T (2012) Policy Depot: a tool to build global capacity in cardiovascular health policy. Global Heart 7:47–51

Mazzati E, Vander Hoorn S, Lopez AD, Danaie G, Rodgers A, Mathers CD, Murray CJL (2006) Comparative quantification of mortality and burden of disease attributable to selected risk factors. In: Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJL (eds) Global burden of disease and risk factors. The International Bank for Reconstruction and Development/The World Bank, Washington, DC, pp 241–396

McCormick J, Skrabanek P (1988) Coronary heart disease is not preventable by population interventions. Lancet 2:839–841

Murray CJL, Lopez AD (eds) (1996) The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020. Harvard School of Public Health, Boston

Narayan KMV, Williams D, Gregg EW, Cowie CC (2011) Diabetes public health: from data to policy. Oxford University Press, New York

National Forum for Heart Disease and Stroke Prevention (2012) www.nationalforum.org. Accessed on 18 Aug 2012

National Heart, Lung, and Blood Institute (2006) Incidence & prevalence: 2006 Chart Book on Cardiovascular and Lung Diseases. US Department of Health and Human Services, Public Health Service, National Institutes of Health, Washington, DC

Olshansky SJ, Ault AB (1986) The fourth stage of the epidemiologic transition: the age of delayed degenerative diseases. Milbank Q 64:355–391

Omran AR (1971) The epidemiological transition: a theory of the epidemiology of population change. Milbank Q 49:509–538

Pathobiological Determinants of Atherosclerosis in Youth (PDAY) Research Group (1990) Relationship of atherosclerosis in young men to serum lipoprotein cholesterol concentrations and smoking. JAMA 264:3018–3024

Pearson TA, Jamison DT, Trejo-Gutierrez J (1993) Cardiovascular disease. In: Jamison DT, Mosley WH, Measham AR, Bobadilla JL (eds) Disease control priorities in developing countries. Oxford University Press, Oxford, pp 577–594

Poulter NR, Khaw KT, Hopwood BEC, Mugambi M, Peart WS, Rose G, Sever PS (1990) The Kenyan Luo migration study: observations on the initiation of a rise in blood pressure. BMJ 300:967–972

President's Council on Physical Fitness and Sports (1996) Physical activity and health: a report of the Surgeon General. National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, US Department of Health and Human Services, Atlanta, GA

Puska P (2002) Successful prevention of non-communicable diseases: 25 year experiences with North Karelia Project in Finland. Public Health Med 4:5–7

Puska P, Tuomilehto J, Nissinen A, Vartiainen E (eds) (1998) The North Karelia project: 20 year results and experiences. National Public Health Institute, KTL, Helsinki

Roger VL, Go AS, Lloyd-Jones DM, Adams RJ, Berry JD, Brown TM, Carnethon MR, Dai S, de Simone G, Ford ES, Fox CS, Fullerton HJ, Gillespie C, Greenlund KJ, Hailpern SM, Heit JA, Ho PM, Howard VJ, Kissela BM, Kittner SJ, Lackland DT, Lichtman JH, Lisabeth LD, Makuc DM, Marcus GM, Marelli A, Matchar DB, McDermott MM, Meigs JB, Moy CS, Mozaffarian D, Mussolino ME, Nichol G, Paynter NP, Rosamond WD, Sorlie PD, Stafford RS, Turan TN, Turner MB, Wong ND, Wylie-Rosett J, on behalf of the American Heart Association Statistics Committee and Stroke Statistics Subcommittee, Roger VL, Turner MB on behalf of the American Heart Association Heart Disease and Stroke Statistics Writing Group (2011) Heart disease and stroke statistics – 2011 update: a report from the American Heart Association. Circulation 123:e18–e209

Rose G (1981) Strategy of prevention: lessons from cardiovascular disease. BMJ 282:1847–1851

Rose G (1985) Sick individuals and sick populations. Int J Epidemiol 14:32–38

Rose GA, Blackburn H (1968) Cardiovascular survey methods. World Health Organization, Geneva

Rose G, Shipley M (1986) Plasma cholesterol concentration and death from coronary heart disease: 10 year results of the Whitehall study. BMJ 293:306–307

Sarti C, Stegmayr B, Tolonen H, Mahonen M, Tuomilehto J, Asplund K (2003) Are changes in mortality from stroke caused by changes in stroke event rates or case fatality? Results from the WHO MONICA project. Stroke 34(8):1833–1840

Sartwell PE (1959) On the methodology of investigations of etiologic factors in chronic diseases. Further comments. J Chron Dis 11:61–63

Sing CF, Haviland MB, Templeton AR, Zerba KE, Reilly SL (1992) Biological complexity and strategies for finding DNA variations responsible for inter-individual variation in risk of a common chronic disease, coronary artery disease. Ann Med 24:539–547

Stallones RA (1980) To advance epidemiology. Annu Rev Public Health 1:69–82

Stamler J (1992) Established major risk factors. In: Marmot M, Elliott P (eds) Coronary heart disease epidemiology: From aetiology to public health. Oxford University Press, Oxford, pp 35–66

Stamler J, Stamler R, Neaton JD, Wentworth W, Daviglus ML, Garside D, Dyer AR, Liu K, Greenland P (1999) Low risk-factor profile and long-term cardiovascular and noncardiovascular mortality and life expectancy. Findings for 5 large cohorts of young adult and middle-aged men and women. JAMA 282:2012–2018

Stehbens WE (1992) Causality in medical science with particular reference to heart disease and atherosclerosis. Perspect Biol Med 36:97–119

Strasser T (1978) Reflections on cardiovascular diseases. Interdiscip Sci Rev 3:225–230

Tejada C, Strong JP, Montenegro MR, Restrepo C, Solberg LA (1968) Distribution of coronary and aortic atherosclerosis by geographic location, race, and sex. Lab Invest 18:509–526

Thom TJ, Epstein FH, Feldman JJ, Leaverton PE, Wolz M (1992) Total mortality and mortality from heart disease, cancer and stroke from 1950 to 1987 in 27 countries: highlights of trends and their interrelationships among causes of death. NIH Publication 92–3088. National Heart, Lung and Blood Institute, National Institutes of Health, Public Health Service, US Department of Health and Human Services, Bethesda, MD

Tobacco Control (2012) 21 (2) March, 2012 (Entire issue)

Tunstall-Pedoe H (ed) (2003) MONICA monograph and multimedia source book. World Health Organization, Geneva

United Nations General Assembly (2011) Political Declaration of the High-level Meeting of the General Assembly on the Prevention and Control of Non-communicable Diseases. Annex to A/66/L.1.United Nations General Assembly. New York, 16 September 2011. www.un.org/en/ga.. Accessed on 1 Oct 2011

US Department of Health and Human Services (2003) A public health action plan to prevent heart disease and stroke. Centers for Disease Control and Prevention, US Department of Health and Human Services, Atlanta, GA

US Department of Health and Human Services (2006) The health consequences of involuntary exposure to tobacco smoke: a report of the Surgeon General. US Department of Health and Human Services, Centers for Disease Control and Prevention, Coordinating Center for Health Promotion, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta, GA

Vartiainen E, Puska P, Pekkanen J, Tuomilehto J, Jousilahti P (1994) Changes in risk factors explain changes in mortality from ischemic heart disease in Finland. BMJ 309:23–27

Wolf-Maier K, Cooper RS, Banegas JR, Giampoli S, Hense H-W, Joffres M, Kastarinen M, Poulter N, Primatesta P, Rodríguez-Artalejo F, Stegmayr B, Thamm M, Tuomilehto J, Vanuzzi D, Vescio F (2003) Hypertension prevalence and blood pressure levels in 6 European countries, Canada, and the United States. JAMA 289:2363–2369

World Health Organization (1992) International statistical classification of diseases and related health problems: tenth revision. World Health Organization, Geneva

World Health Organization (1999) Report by the Director-General: global strategy for the prevention and control of noncommunicable diseases. World Health Organization, Geneva

World Health Organization (2002) The world health report 2002: reducing risks, promoting healthy life. World Health Organization, Geneva. http://www.who.int/whr/2002/en/. Accessed on 18 Aug 2012

World Health Organization (2003) WHO framework convention on tobacco control. World Health Organization, Geneva

World Health Organization (2006) Commission on Social Determinants of Health. WHO/EPI/EQH/ 01/2006. World Health Organization, Geneva

World Health Organization (2007) Reducing salt intake in populations. Report of a WHO Forum and Technical Meeting 5–7 October 2006, Paris, France. World Health Organization, Geneva http://www.who.int/dietphysicalactivity/reducingsalt/. Accessed on 18 Aug 2012

World Health Organization (2008) Report on the global tobacco epidemic, 2008: The MPOWER Package. World Health Organization, Geneva

World Health Organization (2011) Non-communicable diseases country profiles. World Health Organization, Geneva

World Health Organization (2012) International Classification of Diseases (ICD) information sheet. www.who.int/classifications/icd/factsheet/en/. Accessed on 18 Aug 2012

World Health Organization Collaborative Study of Cardiovascular Disease and Steroid Hormone Contraception (1996) Haemorrhagic stroke, overall stroke risk, and combined oral contraceptives: results of an international, multicenter, case-control study. Lancet 348:505–510

World Health Organization Expert Committee (1990) Prevention in childhood and youth of adult cardiovascular diseases: time for action. WHO Technical Report Series 792. World Health Organization, Geneva

World Health Organization MONICA Project Principal Investigators (1988) The World Health Organization MONICA Project (monitoring trends and determinants in cardiovascular disease): a major international collaboration. J Clin Epidemiol 41:105–114

World Health Organization Study Group (1990) Diet, nutrition, and the prevention of chronic diseases. WHO Technical Report Series 797. World Health Organization, Geneva

World Heart Federation (2012) www.worldheart.org. Accessed on 18 Aug 2012

Yach D, Stuckler D, Brownell KD (2006) Epidemiologic and economic consequences of the global epidemics of obesity and diabetes. Nat Med 1:62–66

Yusuf S, Reddy S, Ounpuu S, Anand S (2001) Global burden of cardiovascular diseases. Part I: general considerations, the epidemiologic transition, risk factors, and the impact of urbanization. Circulation 104:2746–2753

Yusuf S, Hawken S, Ôunpuu S, Dans T, Avezum A, Lanas F, McQueen M, Budaj A, Pais P, Varigos J, Liu L on behalf of the INTERHEART Study Investigators (2004) Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet 364:937–952

Cancer Epidemiology

53

Paolo Boffetta

Contents

P. Boffetta
Institute for Translational Epidemiology and Tisch Cancer Institute, Mount Sinai School of Medicine, New York, NY, USA

International Prevention Research Institute, Lyon, France

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_36,

53.1 Introduction

Cancer encompasses a family of several hundreds of diseases which are distinguished by site, morphology, clinical behavior, and response to therapy. Whether considered from a biological, a clinical, or a public health point of view, it is the malignant and invasive nature of many of these diseases and their ability to spread to distant organs (metastasis) that are of dominant importance.

Although the words “cancer” and “carcinoma” refer to malignant tumors arising from epithelial tissues, they are often used to include all malignant neoplasms. These are characterized by progressive and variable growth of tissue with structural and functional changes with respect to the normal tissue. In many cases, the alterations can be so important that it becomes difficult to identify the organ or the tissue of origin.

Knowledge about the causes of and the possible preventive strategies for malignant neoplasms has greatly advanced during the last decades. This was largely due to the findings from cancer epidemiology. In parallel to the identification of the causes of cancer, primary preventive strategies have been developed. Secondary preventive approaches have also been proposed, and, in some cases, they have been shown to be effective. A careful consideration of the achievements of cancer research, however, suggests that the advancements in knowledge about the causes of cancer have not been followed by an equally important reduction in the burden of cancer. Part of this paradox is explained by the long latency occurring between exposure to carcinogens and development of the clinical disease. Changes in exposure to risk factors are not followed immediately by changes in disease occurrence. The main reasons for the gap between knowledge and public health action rest with the cultural, societal, and economic aspects of exposure to carcinogens.

53.2 Scope and Approaches of Cancer Epidemiology

Cancer epidemiology investigates the distribution and determinants of the incidence, mortality, and prevalence of cancer in human populations (Lagiou et al. 2008). Many approaches have been used in cancer epidemiology which can be classified according to different dimensions, as shown in Table 53.1. Although most studies in cancer epidemiology are observational in nature, intervention (experimental) studies are conducted to evaluate the efficacy of prevention strategies, such as screening programs and chemoprevention trials (clinical trials are usually considered to be outside the scope of cancer epidemiology). Observational studies are traditionally classified in descriptive, analytical (or etiological), and ecological studies (for a detailed description of the different types of epidemiological studies, see chapters ►Descriptive Studies, ►Cohort Studies, ►Case-Control Studies, ►Modern Epidemiological Study Designs, ►Intervention Trials, and ►Use of Health Registers of this handbook).

Table 53.1 Approaches used in cancer epidemiology

Dimension	Approaches	Examples
Nature of observation	Experimental	Chemoprevention trial
	Observational	Cohort study
Purpose of investigation	Description	Time-trend analysis
	Etiological research	Case-control study
	Evaluation	Community trial of screening modalities
Unit of observation	Grouped data	Ecological study of environmental exposure
	Individual data	Case-control study with questionnaire data
Sampling strategy[a]	Census-based	Cohort study
	Sample-based	Case-control study
Source of information on exposure	Routine collection	Record-linkage study
	Ad hoc collection	Questionnaire-based study

[a] In studies based on individual data

Descriptive cancer epidemiology is a particularly flourishing branch of the discipline, thanks to the availability of high-quality population-based cancer registries in many areas of the world and to the possibility to use mortality data to estimate the incidence of highly lethal cancers. As an illustration, Fig. 53.1 shows the estimated incidence of cancer among women in all countries of the world: these estimates are derived from cancer registries and mortality statistics. Although subject to various sources of error, such estimates are more precise than those available for any other chronic disease. A useful distinction of etiological studies concerns the nature of the information on exposure: while some studies use data routinely collected for other purposes, such as census records and hospital files, in other circumstances, ad hoc information on exposure is collected following a variety of approaches, including record abstraction, questionnaires, pedigree reconstruction, environmental monitoring, and measurement of biological markers.

Given the importance of cancer in high-income countries and the efforts to prevent it, cancer epidemiology has acquired a recognized status in medicine and has developed into a separate profession. For this reason, and thanks to the availability of high-quality data on the outcomes of interest, it has played an important role in the development of modern epidemiology. The criteria for causal inference in observational research (with the corollary of methodological studies on bias, confounding, and statistical power) have been largely shaped following the discovery of the important role of tobacco smoking as a human carcinogen (Doll 1998), modern statistical approaches such as multivariable logistic and Poisson regressions have originally been proposed for use in cancer studies (Breslow and Day 1980, 1987, see also chapter ▶Regression Methods for Epidemiological Analysis of this handbook); molecular epidemiology has developed as a discipline

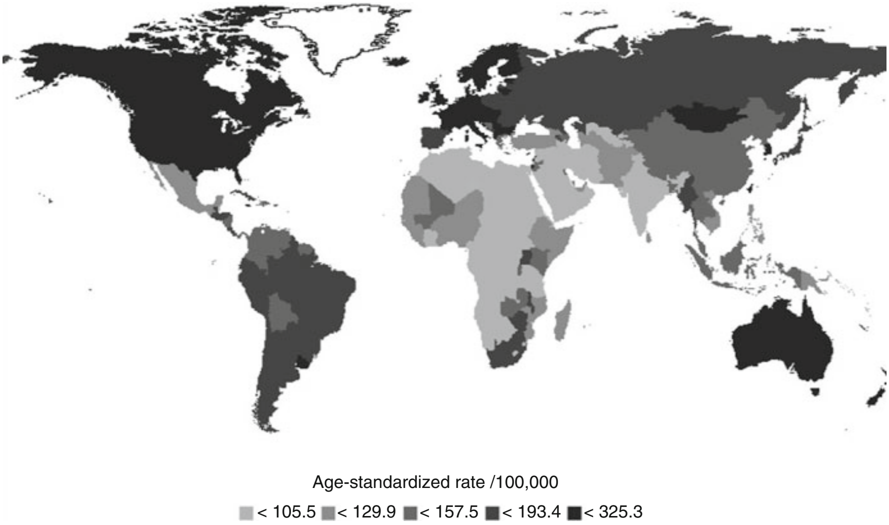


Fig. 53.1 Estimated incidence rate of cancer in women, by country (year 2008) (From Ferlay et al. (2010))

bridging different areas of cancer research (Rothman et al. 2012; see also chapter ▶Molecular Epidemiology of this handbook), and methodological advances in genetic epidemiology have stemmed from familial studies of cancer (Thomas 2000; see also chapter ▶Statistical Methods in Genetic Epidemiology of this handbook).

53.3 The Global Burden of Cancer

The number of new cases of cancer that occurred worldwide in 2008 has been estimated at about 12.7 million (Table 53.2) (Ferlay et al. 2010), including 5.3 million in men and 4.7 million in women. About 5.6 million cases occurred in high-income countries (North America, Japan, Europe including Russia, Australia, and New Zealand) and 7.1 million in middle- and low-income countries. Among men, lung, prostate, colorectal, stomach, and liver cancers are the most common malignant neoplasms (Fig. 53.2), while breast, colorectal, cervical, lung, liver, and ovarian cancers are the most common neoplasms among women (Fig. 53.3).

Such global statistics are of limited interest, given the complexity of the factors affecting the risk of each neoplasm, and the reader is referred to specialized publications for a more detailed review (Jemal et al. 2011). Some general trends can however be identified:

- A decrease in stomach cancer incidence in most countries
- A decrease in the incidence of lung cancer and, to some extent, other tobacco-related cancers among men from high-income countries and a

Table 53.2 Estimated number of new cases of cancer (incidence) and of cancer deaths (mortality) in 2000, by sex and type of country (Ferlay et al. 2010)

	Men	Women	Total
Incidence			
High-income countries	2,964,000	2,591,000	5,555,000
Low-/medium-income countries	3,654,000	3,454,000	7,107,000
Total	6,618,000	6,045,000	12,663,000
Mortality			
High-income countries	1,522,000	1,222,000	2,745,000
Low-/medium-income countries	2,697,000	2,123,000	4,820,000
Total	4,219,000	3,345,000	7,565,000

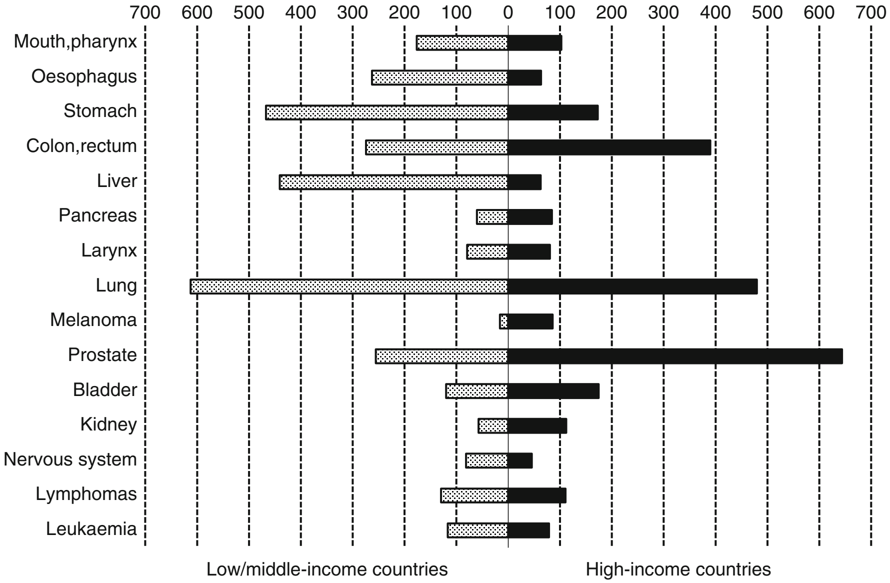


Fig. 53.2 Estimated number of new cancer cases (×1,000) in men (year 2008) (From Ferlay et al. (2010))

corresponding increase among men in middle- and low-income countries and women in high-income countries

- An improvement in survival from low-lethality cancers (e.g., prostate and breast cancer) and a very modest improvement in survival for highly lethal cancers (e.g., esophageal, lung, and liver cancer)

The number of deaths from cancer was estimated at about 7.6 million in 2008 (Table 53.2) (Ferlay et al. 2010). No global estimates of survival from cancer are available: data from selected registries suggest wide disparities between high- and middle-/low-income countries for neoplasms with effective but expensive treatment, such as leukemia, while the gap is narrow for

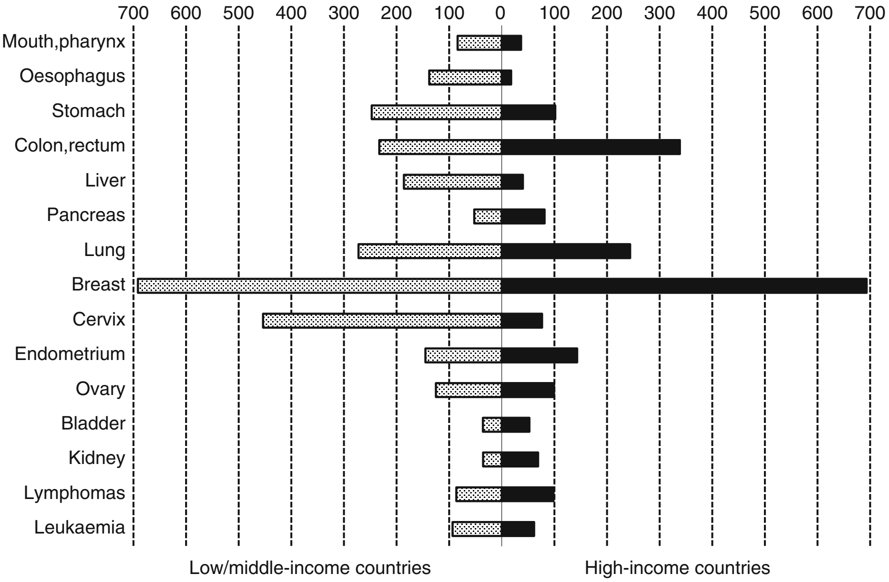


Fig. 53.3 Estimated number of new cancer cases (×1,000) in women (year 2008) (From Ferlay et al. (2010))

neoplasms without an effective therapy, such as lung cancer (Sant et al. 2009; Sankaranarayanan et al. 2011) (Fig. 53.4). The overall 5-year survival of cases diagnosed during 1995–1999 in European Union countries was 50.3% (Sant et al. 2009).

53.4 Causes and Prevention of Human Cancer

In the following sections, the current knowledge about the risk factors and the strategies for primary and secondary prevention of cancer is summarized. For more details, the reader is referred to systematic reviews (Adami et al. 2008; Schottenfeld and Fraumeni 2006).

Several authors have provided estimates of the contribution of known causes to the cancer burden (e.g., Doll and Peto 1981, 2005; HCCP 1996; Boffetta et al. 2009; Parkin et al. 2011; Wang et al. 2012). Such estimates are subject to assumptions and uncertainties and should be interpreted as approximations. In particular, some estimates aimed at attributing all cancers between different causes (e.g., HCCP 1996), while others restricted the exercise to established causes, leaving a substantial fraction of cancers unexplained (e.g., Boffetta et al. 2009; Wang et al. 2012). Table 53.3 shows the results of some of these reviews. It is worth noting that the estimate of the relative importance of the major causes of cancer (e.g., tobacco smoking)

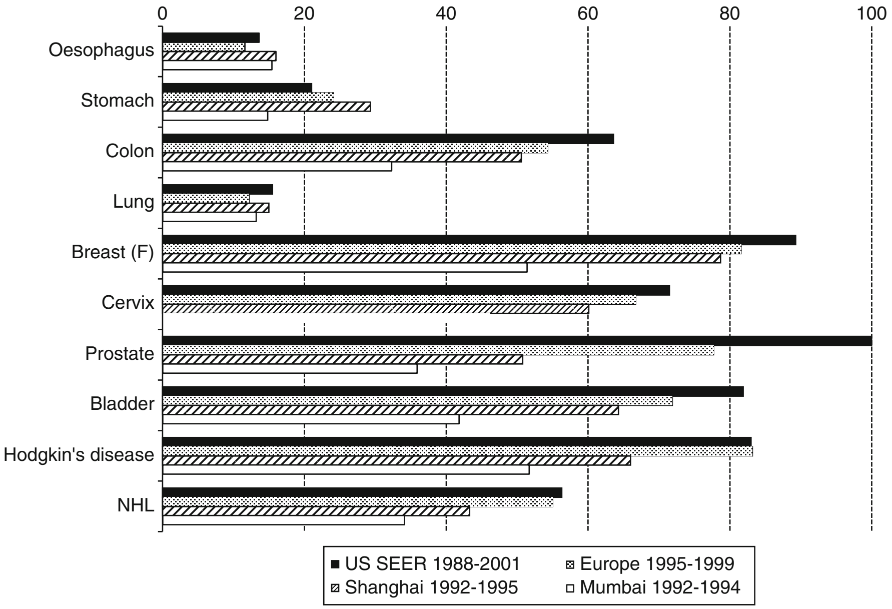


Fig. 53.4 Five-year relative survival (%) from selected cancers (From Ries et al. (2007), Sant et al. (2009), and Sankaranarayanan et al. (2011))

Table 53.3 Selected quantifications of the contribution of major causes to human cancer burden (attributable fractions in percent)

Cause	Boffetta et al. (2009) (France)	Doll and Peto (2005) (UK)	Wang et al. (2012) (China)
Tobacco	24	30	19
Dietary factors	N/A	} 25	7
Obesity	2		1
Sedentary life	2	<1	<1
Biological agents	4	5	27
Occupation	2	2	N/A
Alcohol	7	6	7
Environmental factors	<1	2	N/A
UV/ionizing radiation	1	1	N/A
Reproductive factors	1	15	1
Ionizing radiation	N/A	5	N/A

N/A not available

is fairly consistent. The major role played by infectious agents in China (Wang et al. 2012), as well as in other low- and middle-income countries (de Martel et al. 2012), is an important feature of the important global cancer burden (see Sect. 53.4.7).

53.4.1 Tobacco Smoking

Tobacco smoking is the main single cause of human cancer worldwide. It is a cause of cancers of the oral cavity, pharynx, esophagus, stomach, liver, pancreas, nasal cavity, larynx, lung, cervix, kidney and bladder, and of myeloid leukemia (IARC 2003). It is commonly considered that tobacco smoking causes up to one-third of human cancers (Table 53.3); a detailed review of the number of cancers attributable to tobacco smoking in 1985, which was based on very strict criteria for attribution of cases, resulted in the estimate of at least 15% (Parkin et al. 1994), corresponding to about 1.9 million new cases per year. The estimates were 25% in men and 4% in women, and, in both sexes, they were 16% in developed countries and 10% in developing countries. The low attributable risk in women (and, to a lesser extent, in developing countries) is due to the low consumption of tobacco in past decades: the recent upward trend that has taken place among women and in many developing countries will obviously result in a much greater number of cancers in the future.

The risk of tobacco-related cancers among smokers relative to non-smokers depends on the different characteristics of the habit; in Table 53.4 are reported relative risks found for ever-smokers in Europe and North America. In different populations, risk estimates have been produced for increasing levels of duration and amount of tobacco smoking: in general, a separate effect has been shown for both dimensions of smoking, with a stronger role of the former (USDHHS 2004).

Table 53.4 Relative risk of current smoking and proportion of cancer attributable to tobacco smoking

		Attributable risk (%)[b]	
Cancer	Relative risk for current smoking[a]	Men	Women
Oral cavity	3.4	63	21
Pharynx	6.8	76	47
Esophagus	2.5	51	35
Stomach	1.7	31	15
Liver	1.6	38	18
Pancreas	1.7	25	17
Nasal cavity, nasopharynx	2.0	N/A	N/A
Larynx	7.0	76	65
Lung	9.0	83	70
Cervix	1.8	–	22
Kidney	1.5	26	12
Bladder	2.8	53	37
Myeloid leukemia	1.1	N/A	N/A

N/A not available

[a] Gandini et al. (2008)

[b] Boffetta et al. (2009)

The effect of duration of smoking on the risk of smoking-related cancers, and of lung cancer in particular, is so strong that it is difficult to determine whether there is an independent contribution of other factors, such as age and age at starting smoking. Smoking of filtered cigarettes and cigarettes with reduced tar content results in a lower risk of lung and other cancers than smoking of cigarettes without filter and with high tar content, although by no means, the former products should be seen as “risk-free” (USDHHS 2004). Smoking of black tobacco cigarettes appears to entail a higher risk of most smoking-related cancers than smoking of blond tobacco cigarettes. A carcinogenic effect of cigar and pipe smoking has been demonstrated for cancers of the oral cavity, pharynx, larynx, lung, and bladder. Similarly, smoking of local tobacco products, such as hookah in North Africa and West Asia and bidis in South Asia, entails an increased risk of cancer of the lung and other organs.

A benefit of quitting tobacco smoking in adulthood has been shown for most cancers causally associated with the habit. This result emphasizes the need to devise anti-smoking strategies that address avoidance of the habit among the young as well as reduction of smoking and quitting among adults. There is strong evidence of a protective effect of quitting smoking at any age (Peto et al. 2000). The decline in tobacco consumption that has taken place during the last 20 years among men in North America and several European countries, and which has resulted in decreased incidence of and mortality from lung cancer, has resulted primarily from the increase in the number of smokers quitting at middle age.

With the identification of tobacco as a carcinogen for the lung, the causal nature of an association between a chronic disease and a risk factor was for the first time established beyond doubt, representing an important contribution to the development of epidemiology. The association was replicated in various populations, using different approaches, namely, cohort and case-control studies. This discovery was facilitated by several aspects of tobacco smoking: firstly, it is a potent carcinogen, containing – at high concentrations – several agents acting on different stages of the carcinogenic process; secondly, a sizable group in most populations is composed of heavy smokers, exposing themselves to high doses; and thirdly, exposure is easier to quantify compared to most other agents, since smokers can report with a good degree of precision their present and past consumption.

53.4.2 Use of Smokeless Tobacco Products

There is strong epidemiological evidence that use of smokeless tobacco products is associated with an increased risk of head and neck cancer (Boffetta et al. 2008a). Chewing of tobacco-containing products is particularly prevalent in Southern Asia, where it represents a major cause of cancer of the oral cavity, pharynx, esophagus, and larynx, either alone or in combination with smoking. An increased risk of pancreatic cancer has also been suggested.

53.4.3 Dietary Factors

Despite considerable research efforts in cancer epidemiology, the exact role of dietary factors in causing human cancer remains largely obscure. The World Cancer Research Fund (WCRF 1997, 2007) has published two systematic reviews of the evidence of an association between intake of foods, food groups and nutrients, and different cancers. The second set of evaluations are summarized in Table 53.5. In general, according to WCRF, there are few evaluations of convincing evidence: for fruits and vegetables, the evidence was considered suggestive only for cancers of upper digestive organs and lung. In recent years, the evidence has grown for a carcinogenic role of excess caloric intake, disregarding the source of calories, resulting in overweight and obesity (see Sect. 53.4.4).

The mechanisms of dietary-related carcinogenesis are not well understood. Dietary factors may play a role in most if not all steps of the process, including genotoxicity, interference in the metabolism of other carcinogens, methylation of cancer genes, alteration of DNA repair and apoptotic mechanisms, alteration of DNA and cell replication, and cell proliferation (for a review, see WCRF 2007). In particular, it is plausible that nutrients in fresh fruits and vegetables act at least in part via control of endogenously formed radical oxygen species. In addition, insulin and insulin-like Growth Factor-I (IGF-I), which are linked to caloric intake from diet, may inhibit apoptosis and stimulate cell proliferation (see Sect. 53.4.11).

Several suspected dietary carcinogens have been widely studied in cancer epidemiology. Grilled and barbecued meat and fish contain carcinogenic polycyclic aromatic hydrocarbons and heterocyclic amines: high intake of these foods has been

Table 53.5 Assessment of associations between dietary factors and human cancer (WCRF 2007)

Factor (high intake)	Oral cavity, pharynx, larynx	Esophagus	Stomach	Colon, rectum	Lung	Prostate
Non-starchy vegetables	(−)	(−)	(−)			
Allium vegetables			(−)	(−)		
Fruits	(−)	(−)	(−)		(−)	
Red meat				+		
Processed meat				+		
Maté		(+)				
Salt			(+)			
β-carotene					(+)	
Lycopene						(−)
Selenium						(−)
Calcium				(−)		

Direction of the effect: + increased risk; − decreased risk

Strength of the evidence: *no brackets*: convincing evidence of an association; *brackets*: possible association suggested to increase the risk of stomach and colorectal cancer. Similarly, high intake of cured and processed meat is a probable cause of digestive tract cancer: nitrosamines might be among the relevant carcinogens. High intake of salt probably increases the risk of stomach cancer, and that of fermented fish (e.g., Cantonese-style salted fish) may increase the risk of cancer of the nasopharynx (WCRF 2007).

In several areas of Asia and Africa, high incidence of liver cancer is due to food contamination by mycotoxins, including aflatoxins (WCRF 2007). In Central Europe, a chronic renal disease called Balkan endemic nephropathy has been described, which is associated with an increased risk of kidney cancer and is due to contamination of foodstuff with aristolochic acid, derived from herbs of the *Aristolochia* genus (Grollman et al. 2007). The same agent is responsible for kidney cancer in Taiwan and China, where exposure occurs mainly as result of use of traditional herb-based medicines (Chen et al. 2012).

Intake of large amounts (more than 1 liter per day) of hot maté, a herbal tea, is a possible risk factor for esophageal cancer in Southern Brazil, Uruguay, and Northern Argentina (WCRF 2007). It is unclear, however, whether the effect is due to components of maté or to the high temperature: studies from other regions suggest that intake of hot beverages (e.g., hot tea in Iran, Singapore, and Japan; hot coffee in Puerto Rico; and hot drinks or soups in Hong Kong) increases the risk of esophagitis and esophageal cancer, although the evidence is less consistent than in the case of maté (Nyrén and Adami 2008a).

The investigation of dietary carcinogens presents major challenges because of the difficulties to assess precisely the relevant carcinogenic (or preventive) factors. In most populations, diet varies greatly during the life of an individual, because of changes in personal choices and in societal aspects (availability of different food items, modification of eating patterns, etc.). Furthermore, many nutritional factors are strongly correlated, making it difficult to disentangle the effect of each factor, and variability in exposure within relatively homogeneous populations might not be large enough to allow the detection of carcinogenic effects. Dietary retrospective exposure assessment is complicated by recall bias and unavailability of valid biomarkers, making the case-control approach particularly unsuitable. Even the evidence derived from prospective (e.g., cohort) studies, however, is far from being unequivocal: as an example, the fairly established notion that high intake of fat, mainly of saturated fat from animal foods, might be a risk factor for breast cancer was challenged by the results of prospective studies based on detailed dietary assessment Holmes et al. (1999). In recent years, there is growing interest in the study of the effects of dietary patterns on cancer risk (Verberne et al. 2010). For a general discussion of nutritional epidemiology, see chapter ▸Nutritional Epidemiology of this handbook.

53.4.4 Overweight and Obesity

Overweight, defined as body mass index (BMI) over 25 kilograms per square meter, increases the risk of colon, pancreas, breast (postmenopausal), endometrial, and kidney cancer and of adenocarcinoma of the esophagus (WCRF 2007). The risk of these cancers is linearly related to severity of overweight and obesity, where obesity is defined as BMI over 30 kg/m^2; adult weight gain is a strong and consistent predictor of risk. In the case of colon cancer, body fat distribution expressed as waist-to-hip circumference ratio resulting in abdominal fatness might have an effect independent from that of body mass. It is likely that obesity exerts a carcinogenic effect via alteration of endogenous hormone metabolism, involving in particular insulin resistance and chronic hyperinsulinemia, modulation of adrenal cortical hormones, and increased bioavailability of estrogens. Other possible mechanisms include interference with carcinogen metabolism, accumulation of reactive oxygen species, and alteration of mechanisms regulating cell proliferation, resulting in enhanced proliferation and reduced apoptosis, as well as induction of angiogenesis in tissues other than the fat. The magnitude of the excess risk is not very high (for most cancers, the relative risk ranges between 1.1 and 1.5 for an increase of 5 kg/m^2); however, the attributable risk in high-income countries is large because of the high prevalence of overweight people: estimates for Europe suggest that about 2.5% of all cancers in women and 4.1% in men are attributable to overweight and obesity (Renehan et al. 2008). For the epidemiology of overweight and obesity, see chapter ▸Epidemiology of Obesity of this handbook.

53.4.5 Physical Activity

Regular sustained workplace or recreational physical activity (e.g., at least 30 minutes per day) decreases the risk of colon cancer; a protective effect is also likely for postmenopausal breast cancer and endometrial cancer (WCRF 2007). The magnitude of risk reduction for colon and breast cancer is in the order of 40%, and a dose-response relationship has been shown for both neoplasms. Although regular physical activity contributes to weight control, the epidemiological evidence suggests that two factors also act independently. The mechanisms through which physical activity contributes to cancer prevention are not fully understood, but they may include enhancement of immune function, interference with sex steroids, and insulin and IGF-I pathways (WCRF 2007). For a detailed description of studies on and methods to assess physical activity, see chapter ▸Physical Activity Epidemiology of this handbook.

53.4.6 Alcohol Drinking

There is convincing epidemiological evidence that the consumption of alcoholic beverages increases the risk of cancers of the oral cavity and pharynx, esophagus, liver, colorectum, larynx, and breast (Boffetta and Hashibe 2006). The risks tend to increase with the amount of ethanol drunk, in the absence of any clearly defined threshold below which no effect is evident; an interaction has been shown between alcohol drinking and tobacco smoking. The evidence of differences in carcinogenicity among alcoholic beverages is not conclusive. Alcohol might act as cocarcinogen, enhancing the effect of tobacco and dietary carcinogens; in addition, a direct carcinogenic effect of acetaldehyde, the main metabolite of ethanol, cannot be excluded. The relative risk for breast cancer is in the order of 1.07 for each 10 grams per day increase in alcohol intake; however, this association is of importance because of the apparent lack of a threshold, the large number of women drinking moderate or large amounts of alcohol, and the high incidence of the disease.

The carcinogenic effect of alcohol should be considered in the light of other health effects, notably the increased mortality from chronic digestive diseases and accidents and the reduced mortality from cardiovascular diseases among moderate drinkers (Ronksley et al. 2011). In middle-aged and old people, the benefit on cardiovascular disease is likely to offset the increased cancer risk, up to a level of approximately 20 g/day among men and 10 g/day among women.

53.4.7 Infectious Agents

There is growing epidemiological evidence that chronic infection with some viruses, bacteria, and parasites represents a major cause of human cancer, in particular, in developing countries. A number of infectious agents have been evaluated within the IARC Monograph Programme (Table 53.6), and the evidence of a causal association has been classified as sufficient for several of them. Human papilloma virus (HPV) is detected in almost all cases of cervical cancer: several oncogenic HPV types have been identified, with HPV 16 and 18 being the most prevalent ones. Chronic infection with Hepatitis B virus (HBV) or Hepatitis C virus (HCV) is a major cause of liver cancer worldwide; an interaction has been shown between HBV infection and other causes of liver cancer such as aflatoxin exposure. HCV infection has also been associated with an increased risk of non-Hodgkin lymphoma. Additional carcinogenic viruses include Epstein-Barr virus, a major cause of Hodgkin's disease and of some types of non-Hodgkin lymphoma; human herpes virus 8 (HHV8), which causes Kaposi sarcoma; human immunodeficiency virus I, which causes various types of non-Hodgkin lymphoma; and human T-cell leukemia/lymphoma virus I. In addition, childhood leukemia is likely linked to one or more viruses that have not yet been identified.

Infection with Helicobacter pylori is associated with an approximately sixfold increased risk of non-cardia gastric cancer, after controlling for other risk factors of the disease (Nyrén and Adami 2008b). Unplanned control of Helicobacter infection via widespread antibiotic use and improved living conditions may have contributed to the decline in stomach cancer incidence, which occurred in many countries during recent decades. Infestation with several parasites has been linked with occurrence of human cancer in tropical countries: the evidence is particularly strong for Schistosoma haematobium, causing bladder cancer in North Africa and the Middle East, and Clonorchis sinensis, causing cholangiocarcinoma in Southeast Asia (IARC 2010).

Table 53.6 Assessment of associations between infections and human cancer (IARC 2010)

	Evidence[a]	Target organs[b]
Viruses		
Hepatitis B virus	S	Liver
Hepatitis C virus	S	Liver, lymphoma
Hepatitis D virus	I	Liver
Human papilloma virus types 16, 18, 31, 33, 35, 39, 45, 51, 52, 58, 59	S	Cervix, anus, penis, oral cavity
Human papilloma virus type 68	L	(Cervix)
Human papilloma virus types 6, 11, β, γ	I	
Human immunodeficiency virus	S	Kaposi sarcoma, non-Hodgkin lymphoma
Human T-cell lymphotropic virus I	S	Adult T-cell leukemia/lymphoma
Human T-cell lymphotropic virus II	I	
Epstein-Barr virus	S	Burkitt lymphoma, Hodgkin disease, nasopharynx
Human herpes virus 8	S	Kaposi sarcoma
Bacterium		
Helicobacter pylori	S	Stomach cancer, gastric lymphoma
Parasites		
Schistosoma haematobium	S	Bladder
Schistosoma japonicum	L	(Liver, stomach)
Schistosoma mansoni	I	
Opisthorchis viverrini	S	Liver
Opisthorchis felineus	I	
Clonorchis sinensis	S	Liver

[a] *I* inadequate, *L* limited, *S* sufficient
[b] Established target organs *without brackets*; suspected target organs *in brackets*

Global estimates of the number of cases of cancer attributable to biological agents suggest that at least 16% of all neoplasms worldwide are due to infection (Table 53.7) (de Martel et al. 2012). HBV- and HCV-related liver cancer, HPV-related cervical cancer, and Helicobacter-related stomach cancer each account for approximately 30% of the total. Because of the high prevalence of most carcinogenic agents in LMIC, the estimate of the attributable risk is higher in this part of the world (23%).

More than for other causes of cancer, a carcinogenic role of infectious agents is strongly suggested by extreme variability in cancer risk observed among populations in descriptive epidemiological studies. Thus, an infectious agent had been suspected for a long time for a number of human neoplasms (e.g., Kaposi sarcoma), before sensitive and specific assays became available for the identification of the

Table 53.7 Cancer risk attributable to infectious agents (de Martel et al. 2012)

Cancer	Agent	*AF*%	*N* cases
Liver	HBV, HCV	77	580,000
Non-Hodgkin lymphoma	HCV	8	29,000
Stomach, non-cardia	H. pylori	75	650,000
Non-Hodgkin lymphoma, gastric	H. pylori	74	13,000
Cervix	HPV	100	530,000
Vulva	HPV	43	12,000
Vagina	HPV	70	9,000
Penis	HPV	50	11,000
Anus	HPV	88	24,000
Oropharynx	HPV	26	22,000
Nasopharynx	EBV	85	72,000
Hodgkin lymphoma	EBV	49	33,000
Burkitt lymphoma	EBV	63	7,000
Adult T-cell leukemia	HTLV-I	100	2,000
Kaposi sarcoma	HHV8	100	43,000
Bladder	S. haematobium	2	6,000
Total		16	2,000,000

AF attributable fraction

responsible agent. In addition, the investigation of infectious causes of cancer poses special problems of reverse causality: the detection of an agent in a tumor, as compared to the normal tissue of the patients or controls, does not imply an etiological role, since the altered environment resulting from the neoplasm might favor the growth of the microorganism above detection levels. Cohort studies with repeated samples of the target tissue or surrogate material (typically serum) represent the strongest approach to establish causality.

53.4.8 Occupational Exposures

Several occupational agents, groups of agents and mixtures, as well as exposure circumstances, are classified as carcinogenic by IARC (Table 53.8) (IARC 2012a, b). The evidence linking some of the agents (e.g., beryllium, formaldehyde) to cancer is relatively weak, and others (e.g., mustard gas) are mainly of historical interest; however, exposure is still widespread for important carcinogens such as asbestos, coal tar, and other mixtures of polycyclic aromatic hydrocarbons, heavy metals, and silica. Although the overall burden of occupational cancer is relatively small, these cancers concentrate among exposed subjects (mainly male blue-collar workers), among whom they may represent a sizeable proportion of total cancers (Boffetta et al. 1995). Furthermore, unlike lifestyle factors, exposure is involuntary and can be, to a large extent, avoided. In fact, reduction of exposure to occupational and environmental carcinogens has taken place in industrialized countries during recent

Table 53.8 Occupational agents, classified by IARC as carcinogenic to humans

Agents, groups of agents	*Main industry, use*
4-Aminobiphenyl	Pigment
Arsenic and arsenic compounds	Glass, metal, pesticide
Asbestos	Insulation, filter, textile
Benzene	Chemical, solvent
Benzidine, dyes metabolized to benzidine	Pigment
Benzo[a]pyrene	Combustion fumes
Beryllium and beryllium compounds	Aerospace
Bis(chloromethyl)ether and chloromethyl methyl ether[a]	Chemical intermediate
1,3 Butadiene	Plastics, rubber
Cadmium and cadmium compounds	Dye/pigment
Chromium[VI] compounds	Metal plating, dye/pigment
Dioxin	Chemical
Ethylene oxide	Sterilant
Formaldehyde	Plastics, textiles, laboratory agent
Mustard gas[a]	War gas
2-Naphthylamine	Pigment
Nickel compounds	Metallurgy, alloy, catalyst
PCB-126	Chemical
Plutonium-239 and its decay products	Nuclear industry
Radium-226 and its decay products[a]	Luminizing industry
Radium-228 and its decay products[a]	Luminizing industry
Radon-222 and its decay products	Mining
Silica, crystalline	Stone cutting, mining, glass
Solar radiation	Agriculture
Talc containing asbestiform fibers	Paper, paints
Trichloroethylene	Solvent, dry cleaning, metal
Vinyl chloride	Plastics
X- and γ-radiation	Medical
Mixtures	
Coal-tar pitches	Construction, electrode
Coal tars	Fuel, construction, chemical
Diesel engine exhaust	Transport
Leather dust	Leather
Mineral oils, untreated	Metal
Shale oils	Fuel
Soot	Pigment
Wood dust	Wood
Exposure circumstances	
Aluminum production	
Auramine, manufacture of[a]	Pigment
Boot and shoe manufacture and repair	
Coal gasification	
Coal-tar distillation	

(*continued*)

Table 53.8 (continued)

Exposure circumstances	*Main industry, use*
Coke production	
Hematite mining (underground) with exposure to radon	
Iron and steel founding	
Magenta production[a]	Pigment
Painter (occupational exposure as a)	
Rubber industry	
Strong inorganic-acid mists containing sulfuric acid	Metallurgy

[a] Agent mainly of historical interest

decades and represents one of the successes of cancer epidemiology. Exposures in LMIC, however, might still entail an increased risk of cancer, in particular, in the informal sector where control measures are less readily implemented (Santana and Ribeiro 2011).

The epidemiological approach to study occupational causes of cancer was traditionally based on the historical cohort design. Groups of workers were identified via company or union records, and their cancer mortality or incidence was compared with that of a reference population, most commonly that of the country or the region, leading to the estimate of indirectly standardized mortality (or incidence) ratios. The main reasons for the success of the application of epidemiology to the field of occupational cancer are the possibility to identify clearly defined groups of exposed individuals and the availability of historical measures of exposure. For more details on occupational epidemiology, see chapter ▸Occupational Epidemiology of this handbook.

53.4.9 Environmental Agents

Exposure to many occupational carcinogens listed in Table 53.8 also occurs in the general environment; for two additional agents, the naturally occurring fiber erionite and short-lived radioiodine isotopes, the main source of exposure is the general environment. Overall, the available evidence suggests, in most populations, a small role of purely environmental sources of exposure to carcinogens (air, water, soil pollution): global estimates are in the order of 1% or less of total cancers. This is in contrast with public perception, which often identifies environmental pollution as a major cause of human cancer. It should be stressed, however, that in selected areas (e.g., residence near asbestos-processing plants or in areas with drinking water contaminated by arsenic), environmental exposure to carcinogens may represent an important cancer hazard (Boffetta 2006).

The search for environmental causes of cancer has been particularly elusive to the epidemiological approach. The main reason for such relative lack of success lies in several biases affecting the assessment of exposure to most environmental carcinogens and leading to false-negative results: low-level exposure is often widespread and the range of dose is limited, exposure levels vary with time and most available measurements refer to the present or recent past, and individuals are unable to validly and precisely reconstruct their past exposure. For more details on these problems, please refer to chapter ▸Environmental Epidemiology of this handbook.

53.4.10 Reproductive Factors

The epidemiological evidence of a carcinogenic effect of reproductive factors is strongest for breast cancer: early age at menarche, low parity, late age at first pregnancy, and late age at menopause are all associated with an increased risk, while spontaneous and induced abortions are not (Hankinson et al. 2008). In addition, breastfeeding protects from breast cancer. A large pooled analysis resulted in an estimated 4.3% (95% confidence interval (CI) 2.9–5.8) decrease in risk for every 12 months of breastfeeding, in addition to a decrease of 7.0% (95% CI 5.0–9.0) for each birth (CGHFBC 2002). The same reproductive factors seem to exert an effect on endometrial cancer risk similar to that played on breast cancer, while the evidence of an effect on other cancers is inadequate, although there is limited evidence that nulliparity increases the risk of ovarian cancer. No detailed estimates are available on the contribution of reproductive factors to the global burden of cancer, and the range proposed by different authors vary greatly (from 1% to 15%), depending on the underlying assumptions (Boffetta et al. 2009; Doll and Peto 2005). An extensive discussion of methodological problems in reproductive epidemiology can be found in chapter ▸Reproductive Epidemiology of this handbook.

53.4.11 Hormones

Increased levels of endogenous estrogens are associated with an increased risk of breast and endometrial cancers, and a similar effect is likely to be played by endogenous androgens (Hankinson et al. 2008; DeVivo et al. 2008). The role of other hormones, such as progesterone and prolactin, in these cancers is not clearly known, nor is the role of endogenous androgens in prostate cancer.

There is a large body of evidence that growth hormones, in particular, insulin and IGF-I, have a strong effect on the risk of breast, colon, pancreas, endometrium, prostate, and possibly other cancers (Faulds and Dahlman-Wright 2012). An effect of other endogenous hormones such as growth hormone on cancer risk is plausible but not fully elucidated in epidemiological studies (Clayton et al. 2011). There is a large body of epidemiological studies on cancer risk following exposure to exogenous hormones. Current and recent (up to 10 years) use of oral contraceptives entails a small increase in breast cancer risk, but no excess risk is apparent 10 or more years after cessation of use (CGHFBC 1996). Long-term use of oral contraceptives is associated with an increased risk of liver cancer, while the risk of endometrial and ovarian cancer is decreased following oral contraceptive use (IARC 1999).

Postmenopausal hormonal therapy increases the risk of breast and endometrial cancer (CGHFBC 1997). In the case of breast cancer, the effect is stronger for combined estrogen-progestagen combinations than for other types of hormonal therapy (Beral 2003). The evidence for other organs is inconclusive.

Tamoxifen is widely used for treatment of breast cancer: beyond its therapeutic effects, it decreases the risk of contralateral breast cancer, but it increases the risk of endometrial cancer (Hankinson et al. 2008; DeVivo et al. 2008).

53.4.12 Perinatal Factors

Excess energy intake early in life is possibly associated with breast and colon cancer (WCRF 2007). The role of attained height, growth factors, and other factors such as insulin resistance or sensitivity in this association is unclear. In addition, high birth weight is possibly linked with an increased risk of breast cancer (Hankinson et al. 2008). Perinatal factors have been proposed to cause a sizable proportion of human cancers, but the estimate is subject to uncertainty. The implications of these findings for preventive strategies will be clarified by a more detailed understanding of the underlying carcinogenic mechanisms (Xue and Michels 2007).

53.4.13 Ionizing and Non-Ionizing Radiation

The available epidemiological studies of populations exposed to ionizing radiation following military actions, accidents, occupational exposure, and medical treatments represent a very comprehensive database, which has been used beyond the assessment of radiation carcinogenicity, notably to elaborate models of carcinogenesis in humans and of quantitative risk assessment (Moolgavkar et al. 1999). Ionizing radiation causes acute lymphoblastic leukemia, acute myeloid leukemia, chronic myeloid leukemia, and breast, lung, and thyroid cancers (IARC 2012c). Bone, rectal, and brain cancers may develop following prolonged therapeutic exposure. There is evidence of a linear dose-response relationship between radiation dose and cancer risk. However, levels at which people are commonly exposed to man-made radiation in most countries carry little risk, and the main exposure comes from natural radiation, including indoor radon (UNSCEAR 2010). The estimates of the global contribution of ionizing radiation to human cancer range from 1% to 5% (Table 53.3).

The study of cancer risk following exposure to ionizing radiation represented one of the main paradigms of chronic disease epidemiology. In most exposure circumstances, doses – including those in the past – are known with great precision.

In addition, they are characterized by different intensity and dose rates, allowing the separate investigation of different components of the carcinogenic effect.

Solar (ultraviolet) radiation is carcinogenic to the skin and the lip, and it might increase the risk of other neoplasms such as non-Hodgkin lymphoma (IARC 2012c). Over 90% of skin neoplasms are attributable to sunlight; however, because of the low fatality of non-melanocytic skin cancer, solar radiation is responsible for only up to 1% of total cancer deaths (Table 53.3). Epidemiological studies have contributed to elucidate the contribution of dose rate and time of exposure in ultraviolet-related carcinogenesis. There is no consistent evidence of a carcinogenic effect of other types of non-ionizing radiation, in particular electric and magnetic fields (Feychtin et al. 2005). For more details on radiation epidemiology, please refer to chapter ▸Radiation Epidemiology of this handbook.

53.4.14 Medical Procedures and Drugs

In addition to postmenopausal hormonal therapy, oral contraceptives, and tamoxifen, other drugs may cause cancer. Many cancer chemotherapy drugs are active on the DNA, in order to block the replication of cancer cells. This, however, might result in damage to normal cells, including cancer transformation. The main neoplasm associated with chemotherapy treatment is leukemia, although the risk of solid tumors is also increased. A second group of carcinogenic drugs includes immunosuppressive agents, which have been studied in particular in transplanted patients (Melbye et al. 2008). Non-Hodgkin lymphoma is the main neoplasm caused by these drugs. Phenacetin-containing analgesics increase the risk of cancer of the renal pelvis (IARC 2011).

There is strong evidence from observational studies that aspirin reduces the risk of colorectal cancer (Potter and Hunter 2008), an effect probably shared by other non-steroidal anti-inflammatory drugs.

No precise estimates are available for the global contribution of drug use to human cancer. It is unlikely, however, that drugs represent more than 1% in developed countries (Doll and Peto 2005). Furthermore, the benefits of such therapies are usually much greater than the potential cancer risk.

Use of ionizing radiation for diagnostic purposes is likely to carry a small risk of cancer, which has been demonstrated only for childhood leukemia following intrauterine exposure (IARC 2000). Radiotherapy increases the risk of cancer in the irradiated organs. There is no clear evidence of an increased cancer risk following other medical procedures, including surgical implants.

The epidemiological investigation of the carcinogenicity of drugs and medical procedures shares several characteristics of occupational cancer research: well-defined groups of exposed individuals and valid records of exposure, often in the form of prescription or hospital discharge databases, in addition to the strong potency of several medical agents. These factors explain the relatively large number of drugs identified as human carcinogens.

53.4.15 Medical Conditions

Changes in immunological function are likely to play an important role in human cancer, but epidemiological studies have been largely unable to identify specific factors determining an increased or a decreased risk. Both severe immunosuppression and immunostimulation are associated with an elevated risk of cancer (Morgan et al. 2006). On the one hand, individuals infected with the human immunodeficiency virus (HIV) and patients undergoing immunosuppressive treatments, such as transplant recipients, are at increased risk of lymphoma and skin cancer (Morgan et al. 2006). On the other hand, patients suffering from systemic autoimmune diseases are also at increased risk of lymphoma and possibly other neoplasms (Morgan et al. 2006). The significance of less severe disturbances of the immunological competence is poorly known.

Several chronic inflammatory conditions represent a risk factor for cancer: the epidemiological evidence is particularly strong in the case of colorectal cancer following inflammatory bowel disease and of lymphoma following chronic infectious diseases such as tuberculosis, malaria, and herpes zoster. In addition, gastroesophageal reflux is an important cause of adenocarcinoma of the lower esophagus, a neoplasm whose incidence is increasing in developed countries (Nyrén and Adami 2008a).

53.4.16 Genetic Factors

The notion that genetic susceptibility plays an important role in human cancer is well established, and early studies have demonstrated an increased risk of several types of cancer in individuals with a familial history of the same or related cancers. Several familial conditions entailing a very high risk of cancer have been identified, such as the Li-Fraumeni syndrome and familial polyposis of the colon (Haiman and Hunter 2008). It is only recently that, thanks to the development of molecular tools in human genetics, specific high-risk cancer genes have been identified. Inherited mutations of such high-penetrance cancer genes increase dramatically the risk of some neoplasms (Table 53.9). However, these are rare conditions in most populations, and the number of cases globally attributable to them is rather small.

A familial aggregation has been shown for most types of cancers, in non-carriers of known high-penetrance genes. This is notably the case for cancers of the breast, colon, prostate, and lung. The relative risk is in the order of 2 to 4 and is higher for cases diagnosed at young age. Although some of the aggregation can be explained by shared risk factors among family members, it is plausible that a true genetic component exists for most human cancers. This takes the form of an increased susceptibility to exogenous carcinogens. The knowledge of low-penetrance genes responsible for such susceptibility is still very limited, although research has currently focused on genes encoding for metabolic enzymes, DNA repair, cell cycle control, and hormone receptors (Haiman and Hunter 2008). Current estimates of the

Table 53.9 Examples of high-penetrance cancer predisposition genes (Haiman and Hunter 2008)

Gene	Syndrome	Main targets
RET	Multiple endocrine neoplasia 2	Medullary thyroid carcinoma, pheochromocytoma, parathyroid adenoma
MET	Familial papillary renal cancer syndrome	Papillary renal cancer
APC	Familial adenomatous polyposis	Colorectal cancer
VHL	von Hippel-Lindau syndrome	Clear cell renal carcinoma, hemangioblastoma, retinal angioma, pheochromocytoma
WT1	Wilms' tumor syndrome	Bilateral Wilms' tumor
RB1	Hereditary retinoblastoma	Retinoblastoma
NF1	Neurofibromatosis 1	Neurofibroma, neurofibrosarcoma, optic glioma
NF2	Neurofibromatosis 2	Vestibular schwannoma, meningioma
p53	Li-Fraumeni syndrome	Sarcoma, leukemia, breast, brain, lung, pancreas and skin cancers, others
p16/DCK4	Hereditary melanoma syndrome	Melanoma
PTCH	Nevoid basal cell carcinoma syndrome	Basal cell carcinoma
MEN1	Multiple endocrine syndrome 1	Tumors of parathyroids, gastrointestinal endocrine tissues and anterior pituitary
BRCA1	Hereditary breast-ovarian cancer syndrome	Breast, ovary, prostate and colon cancer
BRCA2	Hereditary breast-ovarian cancer syndrome	Breast (also male) and ovary cancer
PTEN	Cowden syndrome	Hamartoma, breast and thyroid cancer
hMSH2, hMLH1, hPMS1, hPMS2	Hereditary non-polyposis colon	Colon, endometrium, ovary, stomach cancers, others
ATM	Ataxia-telangiectasia	Leukemia, lymphoma, breast cancer
XP(A-G)	Xeroderma pigmentosum	Skin cancer
BLM	Bloom syndrome	Leukemia, lymphoma, most cancers
FAC, FAA	Fanconi anemia	Acute myeloid leukemia, others
WRN	Werner syndrome	Sarcoma, melanoma, thyroid carcinoma

global contribution of genetic factors to human cancer are in the range of 5% to 10%, of which less than 1% is attributable to high-penetrance genes.

The investigation of high- and medium-penetrance genetic cancer risk factors relies mostly on specific methodological approaches whose discussion goes beyond the scope of this chapter (please refer to chapter ▶Statistical Methods in Genetic Epidemiology of this handbook for more details). In the case of low-penetrance genes, however, association studies have been successful in identifying genetic susceptibility factors. Given the lack of dependence of genetic markers of time and disease development, the case-control approach is particularly suitable for this type of investigation.

53.5 Molecular Markers in Cancer Epidemiology

The fact that the identification of new carcinogens, characterized by complex exposure circumstances and weak effects, has become increasingly difficult with traditional epidemiological approaches, the increasing understanding of mechanisms of carcinogenesis and technical developments in molecular biology and genetics are the main reasons for the great increase of the use of molecular markers in cancer epidemiology (Boffetta 2010; Rothman et al. 2012). While a distinction is made between biomarkers of exposure, intermediate events, disease, outcome, and susceptibility, they are best understood as part of a unique framework, aimed at reducing misclassification in the assessment of exposures, covariates, and outcomes in epidemiological studies. The use of exposure biomarkers in cancer epidemiology aims at measuring the biologically relevant exposure more validly and precisely. In some instances, there is an obvious improvement in using an exposure biomarker, as in the case of urinary markers of aflatoxin and tobacco-specific nitrosamines. Intermediate (effect) biomarkers measure early – in general non-persistent – biological events that take place in the continuum between exposure and cancer development, including cellular or tissue toxicity; chromosomal alterations; changes in DNA, RNA, and protein expression; and alterations in functions relevant to carcinogenesis (e.g., DNA repair, immunological response). The analysis of acquired *TP53* mutations is an example of this potentially important contribution of biomarkers. Biomarkers should be validated, and consideration of sources of bias and confounding in molecular epidemiology studies should be no less stringent than in other types of epidemiological studies (see also chapter ▶Molecular Epidemiology of this handbook).

53.6 Screening for Cancer

Screening is considered to be an effective approach to reduce cancer mortality, because human neoplasms go through several preneoplastic stages before they become biologically relevant and clinically detectable. For most cancers, this process takes years or even decades. The possibility to detect preclinical lesions with the potential to develop to a full cancer is highly appealing and is an area of very active research. The slow evolution of cancer, however, is a strong argument to avoid intervention on lesions that do not have the potential to develop to a full cancer during the lifespan of the individual, in order to avoid undue medical procedures such as surgery or chemotherapy. Furthermore, any screening technique has to be carefully evaluated in terms of efficacy to reduce mortality, compliance, and costs. Carefully conducted trials with mortality as main outcome are needed to demonstrate the effectiveness of screening. In practice, however, the available evidence is often restricted to observational data.

Oral inspection aimed at identifying preneoplastic lesions might be an effective approach for secondary prevention of oral cancer. The inspection can be performed by medically certified professionals, but also, in particular, in high-risk areas in developing countries such as India, by specifically trained health workers. Large-scale preventive trials have provided evidence in favor or against this approach (Sankaranarayanan et al. 2005).

Surveillance via flexible sigmoidoscopy, involving removal of adenomas, is a recommended measure for secondary prevention of colorectal cancer. An additional approach consists of the detection of occult blood in the feces. The method suffers from low specificity and, to a lesser extent, low sensitivity, in particular, in the ability to detect adenomas. However, trials have shown a reduced mortality from colorectal cancer after annual tests, although this is achieved at a high cost due to an elevated number of false-positive cases. Current recommendations for individuals aged 50 and over include either annual fecal occult blood testing or flexible sigmoidoscopy or colonoscopy every 5 years (ECCSGWG 2013).

The most suitable approach for secondary prevention of breast cancer is mammography. The effectiveness of screening by mammography in women older than 50 years has been demonstrated, and programs have been established in various countries (Smith et al. 2010). The effectiveness of mammography in women younger than 50 is not demonstrated. The benefit of other screening approaches, such as physical examination and self-examination, is not known.

Cytological examination of exfoliated cervical cells (the Papanicolaou smear test) is effective in identifying precursor lesions, resulting in a decrease in incidence of and mortality from invasive cervical cancer. The benefit is in the order of a two- to fourfold decreased incidence. There is no conclusive evidence, however, regarding the optimal timing of the test (Miller 1999). Cytological smears are not applicable, however, in countries with limited availability of cytologists and pathologists, and alternative approaches for secondary prevention have therefore been proposed, including visual inspection of the cervix with possible enhancement of precursor lesions by acetic acid (Cuzick et al. 2012). There is substantial evidence base to support HPV testing both in triage of women with equivocal abnormal cytology, in surveillance after treatment of cervical intraepithelial neoplasia (CIN) lesions, and in primary screening of women aged 30 years or older (Arbyn et al. 2012).

Secondary prevention has been proposed for prostate cancer, based on digital rectal examination and measurement of prostate-specific antigen. There is no evidence from controlled trials that either procedure decreases the mortality from prostate cancer (Schröder et al. 2012; Andriole et al. 2012). Despite this lack of evidence, these procedures, in particular the prostate-specific antigen testing, have gained popularity in many countries.

Spiral computerized tomography scanning has been shown to be able to identify small, subclinical lesions in the lung of high-risk individuals (Henschke et al. 1999) and to reduce lung cancer mortality for at least 20% (NLSTRT 2011). For a general discussion of the methodological problems of screening, see chapter ▸Screening of this handbook.

53.7 Conclusions

The application of principles of modern epidemiology to cancer research leads to some methodological considerations of a more general nature. Cancer epidemiology is relatively young, yet it has gained an important status in medicine and is practiced by many professionals around the world. In many respects, cancer epidemiology exemplifies the strengths and the weaknesses of the discipline at large.

On the one hand, cancer epidemiology has the privilege of complete and good-quality disease registries in many populations, covering a broad spectrum of rates and exposures. The network of cancer registries not only provides important clues in terms of etiological and clinical research, for example, via the analysis of geographical and temporal differences in incidence, mortality, and prevalence of different neoplasms, but also allows in many countries the conduct of large-scale, high-quality (and relatively low-cost) record linkage studies (cf. chapter ▸Use of Health Registers of this handbook). Examples of such studies include the analysis of cancer risk in migrants by region of origin and length of stay in the host country, the linkage between census and cancer registry data to assess risk from employment in specific occupations, the analysis of second primary neoplasms in cancer patients, and the risk of cancer following diagnosis of (or hospitalization for) non-neoplastic conditions.

On several occasions, cancer epidemiology has been the key tool to demonstrate the causal role of important cancer risk factors. The best example is the association between tobacco smoking and lung cancer, which led in the early 1960s to the establishment of criteria for causality in observational research (Doll 1998). Other contributions of epidemiology to the elucidation of important causes of human cancer include the demonstration of the role of HPV in cervical cancer, the role of Helicobacter pylori in stomach cancer, and that of solar radiation exposure in skin cancer, as well as the growing body of evidence for a major role of overweight and obesity in the etiology of several important neoplasms. These findings have brought important regulatory and public health initiatives as well as lifestyle changes in many countries of the world. For example, Box 53.1 shows the European Code Against Cancer, which adequately summarizes the current evidence for cancer prevention: these recommendations are mainly based on evidence accumulated via epidemiological studies.

Box 53.1. European Code Against Cancer, Third Version (Boyle et al. 2003)

Many aspects of general health can be improved, and many cancer deaths prevented, if we adopt healthier lifestyles:

1. Do not smoke; if you smoke, stop doing so. If you fail to stop, do not smoke in the presence of non-smokers.
2. Avoid obesity.
3. Undertake some brisk, physical activity every day.
4. Increase your daily intake and variety of vegetables and fruits: eat at least five servings daily. Limit your intake of foods containing fats from animal sources.
5. If you drink alcohol, whether beer, wine, or spirits, moderate your consumption to two drinks per day if you are a man or one drink per day if you are a woman.
6. Care must be taken to avoid excessive sun exposure. It is specifically important to protect children and adolescents. For individuals who have a tendency to burn in the sun, active protective measures must be taken throughout life.
7. Apply strictly regulations aimed at preventing any exposure to known cancer-causing substances. Follow all health and safety instructions on substances which may cause cancer. Follow advice of National Radiation Protection Offices.

There are public health programs that could prevent cancers developing or increase the probability that a cancer may be cured:

8. Women from 25 years of age should participate in cervical screening. This should be within programs with quality control procedures in compliance with *European Guidelines for Quality Assurance in Cervical Screening*.
9. Women from 50 years of age should participate in breast screening. This should be within programs with quality control procedures in compliance with *European Guidelines for Quality Assurance in Mammography Screening*.
10. Men and women from 50 years of age should participate in colorectal screening. This should be within programs with built-in quality assurance procedures.
11. Participate in vaccination programs against Hepatitis B virus infection.

These epidemiological "discoveries" share two important characteristics: they involve potent carcinogens, and methods are available to reduce misclassification of exposure to the risk factor of interest and to major possible confounders. It has therefore been possible to consistently demonstrate an association in different human populations. It should be noted that it is not necessary for the prevalence of exposure to be high (although this obviously has an impact on the population attributable risk): examples are the many occupational exposures and medical treatments for which conclusive evidence of carcinogenicity has been established on the basis of epidemiological studies conducted in small populations of individuals with well-characterized high exposure.

On the other hand, when these conditions are not met, the evidence accumulated from epidemiological studies is typically inconsistent and difficult to interpret (Taubes 1995; Boffetta et al. 2008b). The history of cancer epidemiology presents many examples of premature conclusions, which have not been confirmed by subsequent investigations and have damaged the reputation of the discipline. Misclassification of the relevant exposure (cf. chapters ▸Exposure Assessment and ▸Misclassification of this handbook), uncontrolled confounding (cf. chapters ▸Basic Concepts and ▸Confounding and Interaction), and inadequate statistical power (cf. chapter ▸Sample Size Determination in Epidemiological Studies) are the most common limitations encountered in cancer epidemiology. Two solutions have been proposed to overcome these problems. First, epidemiological studies should be very large in size. Second, the use of biological markers of exposure and early effect has been proposed to reduce exposure misclassification, increase the prevalence of the relevant outcomes, and shed light on the mechanism of action of the carcinogen under study (Boffetta and Trichopoulos 2008). These approaches might offer avenues to continue the successful contribution of cancer epidemiology to medical science.

References

Adami H-O, Hunter D, Trichopoulos D (eds) (2008) Textbook of cancer epidemiology, 2nd edn. Oxford University Press, New York

Andriole GL, Crawford ED, Grubb RL 3rd, Buys SS, Chia D, Church TR, Fouad MN, Isaacs C, Kvale PA, Reding DJ, Weissfeld JL, Yokochi LA, O'Brien B, Ragard LR, Clapp JD, Rathmell JM, Riley TL, Hsing AW, Izmirlian G, Pinsky PF, Kramer BS, Miller AB, Gohagan JK, Prorok PC, PLCO Project Team (2012) Prostate cancer screening in the randomized prostate, lung, colorectal, and ovarian cancer screening trial: mortality results after 13 years of follow-up. J Natl Cancer Inst 104:125–132

Arbyn M, Ronco G, Anttila A, Meijer CJ, Poljak M, Ogilvie G, Koliopoulos G, Naucler P, Sankaranarayanan R, Peto J (2012) Evidence regarding human papilloma virus testing in secondary prevention of cervical cancer. Vaccine 30(Suppl 5):F88–F99

Beral V, Million Women Study Collaborators (2003) Breast cancer and hormone-replacement therapy in the Million Women Study. Lancet 362:419–427

Boffetta P (2006) Human cancer from environmental pollutants: the epidemiological evidence. Mutat Res 608:157–162

Boffetta P (2010) Biomarkers in cancer epidemiology: an integrative approach. Carcinogenesis 31:121–126

Boffetta P, Hashibe M (2006). Alcohol and cancer. Lancet Oncol 7:149–156

Boffetta P, Trichopoulos D (2008) Biomarkers in cancer epidemiology. In: Adami H-O, Hunter D, Trichopoulos D (eds) Textbook of cancer epidemiology, 2nd edn. Oxford University Press, New York, pp 109–126

Boffetta P, Kogevinas M, Simonato L, Wilbourn J, Saracci R (1995) Current perspectives on occupational cancer risks. Int J Occup Environ Health 1:315–325

Boffetta P, Hecht S, Gray N, Gupta P, Straif K (2008a) Smokeless tobacco and cancer. Lancet Oncol 9:667–675

Boffetta P, McLaughlin JK, La Vecchia C, Tarone RE, Lipworth L, Blot WJ (2008b) False-positive results in cancer epidemiology: a plea for epistemological modesty. J Natl Cancer Inst 100: 988–995

Boffetta P, Tubiana M, Hill C, Boniol M, Aurengo A, Masse R, Valleron AJ, Monier R, de Thé G, Boyle P, Autier P (2009) The causes of cancer in France. Ann Oncol 20:550–555

Boyle P, Autier P, Bartelink H, Baselga J, Boffetta P, Burn J, Burns HJG, Christensen L, Denis L, Dicato M, Diehl V, Doll R, Franceschi S, Gillis CR, Gray N, Griciute L, Hackshaw A, Kasler M, Kogevinas M, Kvinnsland S, La Veccia C, Levi F, McVie JG, Maisonneuve P, Martin-Moreno M, Newton Bishop J, Oleari F, Perrin P, Quinn M, Richards M, Ringborg U, Scully C, Siracka E, Storm H, Tubiana M, Tursz T, Veronesi U, Wald N, Weber W, Zaridze DG, Zatonski W, zur Hausen H (2003) European code against cancer and scientific justification: third version. Ann Oncol 14:973–1005

Breslow NE, Day NE (1980) Statistical methods in cancer research, vol I. The analysis of case-control studies. IARC scientific publications no. 32. International Agency for Research on Cancer, Lyon

Breslow NE, Day NE (1987) Statistical methods in cancer research, vol II. The design and analysis of cohort studies. IARC scientific publications no. 82. International Agency for Research on Cancer, Lyon

CGHFBC (Collaborative Group on Hormonal Factors in Breast Cancer) (1996) Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53,297 women with breast cancer and 100,239 women without breast cancer from 54 epidemiological studies. Lancet 347:1713–1727

CGHFBC (Collaborative Group on Hormonal Factors in Breast Cancer) (1997) Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer. Lancet 350:1047–1059

CGHFBC (Collaborative Group on Hormonal Factors in Breast Cancer) (2002) Breast cancer and breast feeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50,302 women with breast cancer and 96,973 women without the disease. Lancet 360:187–195

Chen CH, Dickman KG, Moriya M, Zavadil J, Sidorenko VS, Edwards KL, Gnatenko DV, Wu L, Turesky RJ, Wu XR, Pu YS, Grollman AP (2012) Aristolochic acid-associated urothelial cancer in Taiwan. Proc Natl Acad Sci USA 109:8241–8246

Clayton PE, Banerjee I, Murray PG, Renehan AG (2011) Growth hormone, the insulin-like growth factor axis, insulin and cancer risk. Nat Rev Endocrinol 7:11–24

Cuzick J, Bergeron C, von Knebel Doeberitz M, Gravitt P, Jeronimo J, Lorincz AT, J L M Meijer C, Sankaranarayanan R, J F Snijders P, Szarewski A (2012) New technologies and procedures for cervical cancer screening. Vaccine 30(Suppl 5):F107–F116

de Martel C, Ferlay J, Franceschi S, Vignat J, Bray F, Forman D, Plummer M (2012) Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. Lancet Oncol 13:607–615

DeVivo I, Persson I, Adami HO (2008) Endometrial cancer. In: Adami H-O, Hunter D, Trichopoulos D (eds) Textbook of cancer epidemiology, 2nd edn. Oxford University Press, New York, pp 468–493

Doll R (1998) Uncovering the effects of smoking: historical perspective. Stat Methods Med Res 7:87–117

Doll R, Peto R (1981) The causes of cancer. Oxford University Press, New York, NY

Doll R, Peto R (2005) Epidemiology of cancer. In: Warrell DA, Cox TM, Firth JD (eds) Oxford textbook of medicine, 4th edn. Oxford University Press, Oxford, pp 193–218

ECCSGWG (European Colorectal Cancer Screening Guidelines Working Group) (2013) European guidelines for quality assurance in colorectal cancer screening and diagnosis: overview and introduction to the full supplement publication. Endoscopy 45:51–59

Faulds MH, Dahlman-Wright K (2012). Metabolic diseases and cancer risk. Curr Opin Oncol 24:58–61

Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM (eds) (2010) GLOBOCAN 2008: cancer incidence and mortality worldwide. IARC CancerBase series and related publications. International Agency for Research on Cancer, Lyon. See also http://globocon.iarc.fr. Accessed 3 Jan 2013

Feychting M, Ahlbom A, Kheifets L (2005) EMF and health. Annu Rev Public Health 26:165–189

Gandini S, Botteri E, Iodice S, Boniol M, Lowenfels AB, Maisonneuve P, Boyle P (2008) Tobacco smoking and cancer: a meta-analysis. Int J Cancer 122:155–164

Grollman AP, Shibutani S, Moriya M, Miller F, Wu L, Moll U, Suzuki N, Fernandes A, Rosenquist T, Medverec Z, Jakovina K, Brdar B, Slade N, Turesky RJ, Goodenough AK, Rieger R, Vukelić M, Jelaković B (2007) Aristolochic acid and the etiology of endemic (Balkan) nephropathy. Proc Natl Acad Sci USA 104:12129–12134

Haiman C, Hunter D (2008) Genetic epidemiology of cancer. In: Adami H-O, Hunter D, Trichopoulos D (eds) Textbook of cancer epidemiology, 2nd edn. Oxford University Press, New York, pp 86–106

Hankinson S, Tamini R, Hunter D (2008) Breast cancer. In: Adami H-O, Hunter D, Trichopoulos D (eds) Textbook of cancer epidemiology, 2nd edn. Oxford University Press, New York, pp 403–445

HCCP (Harvard Center for Cancer Prevention) (1996) Harvard report on cancer prevention, vol 1, causes of human cancer. Cancer Causes Control 7:S3–S58

Henschke CI, McCauley DI, Yankelevitz DF, Naidich DP, McGuinness G, Miettinen OS, Libby DM, Pasmantier MW, Koizumi J, Altorki NK, Smith JP (1999) Early Lung Cancer Action Project: overall design and findings from baseline screening. Lancet 354:99–105

Holmes MD, Hunter DJ, Colditz GA, Stampfer MJ, Hankinson SE, Speizer FE, Rosner B, Willett WC (1999) Association of dietary intake of fat and fatty acids with risk of breast cancer. JAMA 281:914–920

IARC (International Agency for Research on Cancer) (1999) Hormonal contraception and post-menopausal hormonal therapy. IARC monographs on the evaluation of carcinogenic risks to humans, vol 72. International Agency for Research on Cancer, Lyon

IARC (International Agency for Research on Cancer) (2000) Ionizing radiation, Part 1, X- and gamma (γ)-radiation and neutrons. IARC monographs on the evaluation of carcinogenic risks to humans, vol 75. International Agency for Research on Cancer, Lyon

IARC (International Agency for Research on Cancer) (2003) Tobacco smoking and involuntary tobacco smoke. IARC monographs on the evaluation of the carcinogenic risk of chemicals to humans, vol 83. International Agency for Research on Cancer, Lyon

IARC (International Agency for Research on Cancer) (2010) A review of human carcinogens. Part B: biological agents. IARC monographs on the evaluation of carcinogenic risks to humans, vol 100. International Agency for Research on Cancer, Lyon

IARC (International Agency for Research on Cancer) (2011) A review of human carcinogens. Part A: pharmaceuticals. IARC monographs on the evaluation of carcinogenic risks to humans, vol 100. International Agency for Research on Cancer, Lyon

IARC (International Agency for Research on Cancer) (2012a) A review of human carcinogens. Part C: arsenic, metals, fibres, and dusts. IARC monographs on the evaluation of carcinogenic risks to humans, vol 100. International Agency for Research on Cancer, Lyon

IARC (International Agency for Research on Cancer) (2012b) A review of human carcinogens. Part F: chemical agents and related occupations. IARC monographs on the evaluation of carcinogenic risks to humans, vol 100. International Agency for Research on Cancer, Lyon

IARC (International Agency for Research on Cancer) (2012c) A review of human carcinogens. Part D: radiation. IARC monographs on the evaluation of carcinogenic risks to humans, vol 100. International Agency for Research on Cancer, Lyon

Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D (2011) Global cancer statistics. CA Cancer J Clin 61:69–90

Lagiou P, Trichopoulos D, Adami HO (2008) Concepts in cancer epidemiology and etiology. In: Adami H-O, Hunter D, Trichopoulos D (eds) Textbook of cancer epidemiology, 2nd edn. Oxford University Press, New York, pp 127–154

Melbye M, Smedby KE, Trichopoulos D (2008) Non-Hodgkin's lymphoma. In: Adami H-O, Hunter D, Trichopoulos D (eds) Textbook of cancer epidemiology, 2nd edn. Oxford University Press, New York, pp 669–693

Miller AB (1999) Cervix cancer. In: Kramer BS, Gohagan JK, Prorok PC (eds) Cancer screening: theory and practice. Marcel Dekker, New York, pp 195–217

Moolgavkar S, Krewski D, Schwarz M (1999) Mechanisms of carcinogenesis and biologically based models for estimation and prediction of risk. In: Moolgavkar S, Krewski D, Zeise L, Cardis E, Møller H (eds) Quantitative estimation and prediction of human cancer risks. IARC scientific publications 131. International Agency for Research on Cancer, Lyon, pp 179–237

Morgan GJ, Linet MS, Rabkin CS (2006) Immunologic factors. In: Schottenfeld D, Fraumeni JF (eds) Cancer epidemiology and prevention, 3rd edn. Oxford University Press, New York, pp 549–561

NLSTRT (National Lung Screening Trial Research Team), Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 365:395–409

Nyrén O, Adami H-O (2008a) Esophageal cancer. In: Adami H-O, Hunter D, Trichopoulos D (eds) Textbook of cancer epidemiology, 2nd edn. Oxford University Press, New York, pp 196–238

Nyrén O, Adami H-O (2008b) Stomach cancer. In: Adami H-O, Hunter D, Trichopoulos D (eds) Textbook of cancer epidemiology, 2nd edn. Oxford University Press, New York, pp 239–276

Parkin DM, Pisani P, Lopez AD, Masuyer E (1994) At least one in seven cases of cancer is caused by smoking. Global estimates for 1985. Int J Cancer 59:494–504

Parkin DM, Boyd L, Walker LC (2011) The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. Br J Cancer 105(Suppl 2):S77–S81

Peto R, Darby S, Deo H, Silcocks P, Whitley E, Doll R (2000) Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. BMJ 321:323–329

Potter JD, Hunter D (2008) Colorectal cancer. In: Adami H-O, Hunter D, Trichopoulos D (eds) Textbook of cancer epidemiology, 2nd edn. Oxford University Press, New York, pp 275–307

Renehan AG, Tyson M, Egger M, Heller RF, Zwahlen M (2008) Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. Lancet 371:569–578

Ries LAG, Young JL, Keel GE, Eisner MP, Lin YD, Horner MJ (eds) (2007) SEER Survival Monograph: cancer survival among adults: U.S. SEER Program, 1988–2001, Patient and Tumor Characteristics. National Cancer Institute/SEER Program, Bethesda

Ronksley PE, Brien SE, Turner BJ, Mukamal KJ, Ghali WA (2011) Association of alcohol consumption with selected cardiovascular disease outcomes: a systematic review and meta-analysis. BMJ 342:d671

Rothman N, Hainaut P, Schulte P, Smith M, Boffetta P, Perera F (2012) Molecular epidemiology: principles and practices. IARC scientific publications no. 163. International Agency for Research on Cancer, Lyon

Sankaranarayanan R, Ramadas K, Thomas G, Muwonge R, Thara S, Mathew B, Rajan B, Trivandrum Oral Cancer Screening Study Group (2005) Effect of screening on oral cancer mortality in Kerala, India: a cluster-randomised controlled trial. Lancet 365:1927–1933

Sankaranarayanan R, Swaminathan R, Lucas E (2011) Cancer survival in Africa, Asia, the Caribbean and Central America (SurvCan). IARC scientific publications 162. International Agency for Research on Cancer, Lyon

Sant M, Allemani C, Santaquilani M, Knijn A, Marchesi F, Capocaccia R, EUROCARE Working Group (2009) EUROCARE-4. Survival of cancer patients diagnosed in 1995–1999. Results and commentary. Eur J Cancer 45:931–991

Santana VS, Ribeiro FS (2011) Occupational cancer burden in developing countries and the problem of informal workers. Environ Health 10(Suppl 1):S10

Schottenfeld D, Fraumeni JF (eds) (2006) Cancer epidemiology and prevention, 3rd edn. Oxford University Press, New York

Schröder FH, Hugosson J, Roobol MJ, Tammela TL, Ciatto S, Nelen V, Kwiatkowski M, Lujan M, Lilja H, Zappa M, Denis LJ, Recker F, Páez A, Määttänen L, Bangma CH, Aus G, Carlsson S, Villers A, Rebillard X, van der Kwast T, Kujala PM, Blijenberg BG, Stenman UH, Huber A, Taari K, Hakama M, Moss SM, de Koning HJ, Auvinen A, ERSPC Investigators (2012) Prostate-cancer mortality at 11 years of follow-up. N Engl J Med 366:981–990

Smith RA, Cokkinides V, Brooks D, Saslow D, Brawley OW (2010) Cancer screening in the United States, 2010: a review of current American Cancer Society guidelines and issues in cancer screening. CA Cancer J Clin 60:99–119

Taubes G (1995) Epidemiology faces its limits. Science 269:164–169

Thomas DC (2000) Genetic epidemiology with a capital "E". Genet Epidemiol 19:289–300

UNSCEAR (United Nations Scientific Committee on the Effects of Atomic Radiation) (2010) Sources and effects of ionizing radiation. UNSCEAR 2008 report to the General Assembly with scientific annexes. United Nations, New York

USDHHS (US Department of Health and Human Services), Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health (2004) The health consequences of smoking: a report of the Surgeon General. USDHHS, Washington, DC

Verberne L, Bach-Faig A, Buckland G, Serra-Majem L (2010) Association between the Mediterranean diet and cancer risk: a review of observational studies. Nutr Cancer 62:860–870

Wang JB, Jiang Y, Liang H, Li P, Xiao HJ, Ji J, Xiang W, Shi JF, Fan YG, Li L, Wang D, Deng SS, Chen WQ, Wei WQ, Qiao YL, Boffetta P (2012) Attributable causes of cancer in China. Ann Oncol 23:2983–2989

WCRF (World Cancer Research Fund/American Institute for Cancer Research) (1997) Food, nutrition and the prevention of cancer: a global perspective. World Cancer Research Fund & American Institute for Cancer Research, Washington, DC

WCRF (World Cancer Research Fund/American Institute for Cancer Research) (2007) Food, nutrition, physical activity and the prevention of cancer: a global perspective. World Cancer Research Fund & American Institute for Cancer Research, Washington, DC

Xue F, Michels KB (2007) Intrauterine factors and risk of breast cancer: a systematic review and meta-analysis of current evidence. Lancet Oncol 8:1088–1100

Musculoskeletal Disorders

54

Hilkka Riihimäki

Contents

H. Riihimäki
Centre of Expertise for Health and Work Ability (retired), Finnish Institute of Occupational Health (FIOH), Espoo, Finland

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_37,

54.1 Introduction

Musculoskeletal disorders are a significant public health problem in industrialized countries. They are one of the leading causes of short- and long-term disability and cause high costs due to loss of productivity, burden on health care services, and social security systems. It has been estimated that the cost of back pain only is 1–2% of the gross national product. About 10% is due to direct health care costs and 90% due to indirect costs (Norlund and Waddell 2000). In this chapter, the focus is in morbidity and etiology of musculoskeletal disorders. The social consequences, such as disability, are not discussed, because disability-related issues depend on the societal context, such as cultural factors and social security regulations, to the extent that cross-national comparisons cannot be made. Firstly, an overview is given of the occurrence and risk factors of the most common musculoskeletal disorders, back and neck disorders, osteoarthritis, upper limb disorders, and osteoporosis. Secondly, some methodological problems in epidemiological research, characteristic for musculoskeletal disorders, are discussed including general study design, and assessment of both health outcome and exposure.

54.2 Occurrence and Risk Factors

54.2.1 Back Disorders

Dorsopathies or back disorders are classified in the International Classification of Diseases 10 (ICD-10) (World Health Organisation, WHO 1992) into deforming dorsopathies (kyphosis and lordosis, scoliosis, spinal osteochondrosis, and others), spondylopathies (ankylosing spondylitis, other inflammatory spondyloses, spondylosis, others), and other dorsopathies (intervertebral disc disorders, other dorsopathies not classified elsewhere, and dorsalgia with several subcategories). This classification is not very useful for epidemiological research. For acute back pain, it has been estimated that for 70% of the cases a specific diagnosis cannot be determined, i.e., most of the cases fall into the category dorsalgia in the ICD-10. A large proportion of the societal burden due to back disorders is, however, attributable to non-specific back pain. The origin of such a pain can be in any tissue of the back with pain receptors, but in most cases the origin cannot be determined. An underlying mechanism in back pain is disc degeneration, and the relationship between low back pain and disc degeneration continues to be controversial. Previous pain experience and psychological factors are known to influence pain perception.

Occurrence The prevalence estimates of low back pain (LBP) vary widely from one study to another depending on the population characteristics and assessment methods. It also seems that cultural differences between populations influence perceiving or reporting of back pain (Raspe et al. 2004). In a systematic literature review (Walker 2000), the point prevalence in the general population varied from 12% to 33%, 1-year prevalence from 22% to 65% (Table 54.1), and lifetime

Table 54.1 One-year prevalence of low back pain in community-based good-quality studies (Data from Walker (2000))

Author	Country	Sample size	Age (years)	Prevalence (%)
Harreby et al. (1996)	Denmark	481	38	60/65[a]
Rafnsson et al. (1989)	Iceland	672	16–65	55
Leboeuf-Yde et al. (1996)	Denmark	1,370	30–50	54
Biering-Sorensen (1982)	Denmark	928	30, 40, 50, 60	45
Hillman et al. (1996)	England	3,184	25–64	39
Mason (1994)	England	6,000	16+	37
Walsh et al. (1992)	England	2,667	20–59	36
Lau et al. (1995)	Hong Kong	652	18+	22

[a]men/women

prevalence from 30% to 84% in studies with acceptable quality. A systematic review of LBP prevalence in Africa reported the average 1-year prevalence of 50% among adults (Louw et al. 2007). The prevalence of back pain increases with increasing age, but many studies have shown that after reaching a peak in the middle of the sixth decade, the prevalence declines. A systematic review by Dionne et al. (2006), however, showed that such a curvilinear association between age and back pain prevalence is found only for benign and mixed problems; most studies with severe back pain as the outcome detected an increase of the prevalence with increasing age. Back pain is reported more often by women than men.

A subgroup of back pain, sciatic pain or low back pain radiating to the leg, tends to have a poorer outcome than local low back pain. The prevalence estimates vary considerably from 1.2% to 43% (Konstantinou and Dunn 2008). This is to a great extent due to differences in definitions of the outcome, methods of data collection, and study populations. In only a few studies, clinical assessment has been employed. One example is the Mini-Finland Health Survey, in which a representative sample of 3,322 Finnish men and 3,895 women aged 30 years or more were examined. The point prevalence estimates of clinically verified sciatica or herniated disc were 5.1% in men and 3.7% in women (age-adjusted). For low back pain syndrome, the estimates were 17.5% and 16.3%, respectively (Heliövaara et al. 1991, 1993).

Little data are available on the incidence of low back pain. One reason for this is the episodic nature of back pain which makes it difficult to define the first-time event, and therefore, usually only the incidence of new episodes can be assessed. Biering-Sorensen (1982) reported 6% 1-year cumulative incidence in a Danish community sample aged 20, 30, 40, and 50 years, and Hillman et al. (1996) observed 4.7% annual incidence in an English community sample aged 25–64 years. In an occupational cohort, the 3-year cumulative incidence of LBP was 26.6% (Hoogendoorn et al. 2000a). In a cohort of newly employed workers without LBP at the baseline, new-onset LBP was reported by 19% at 12 and 24 months (Harkness et al. 2003). Most prospective studies among adults have shown no marked influence of age on the incidence of LPB.

Low back pain is common also among children and adolescents. The prevalence seems to reach the level in adults by around 18 years of age. According to a systematic review by Jeffries et al. (2007), the lifetime prevalence of LBP in different populations aged between 7 and 21 years ranged from 7% to 72%. In most cases, symptoms are mild, non-specific, and self-limiting (Blyth et al. 2009). The prevalence is higher among girls than boys, and there is an increasing trend from lower to upper social classes. Low back pain in adolescence seems to predispose to LBP in adulthood, although only a few prospective studies exist. In an 8-year follow-up of nearly 10,000 Danish 12- to 22-year-old twins, those who had persistent LBP at baseline had fourfold risk of reporting LBP during the last year of follow-up as compared with those without LBP (Hestbaek et al. 2006).

A common belief is that the prevalence of back pain has been increasing during the past decades, but the available evidence is somewhat contradictory.

In Finland, 14-year surveillance with annually repeated postal questionnaire surveys among random population samples of 15- to 64-year-old Finns showed no significant change from 1978 to 1992 (Leino et al. 1994). Since then, the surveillance has been continued and, according to unpublished data, the prevalence level has remained rather stable (1-month prevalence of back pain about 35%, 1-year prevalence of back disorder verified by a doctor about 15%). No increase of the 1-month prevalence was obtained either in another 10-year surveillance program (surveyed every fifth year) among random population samples of 30- to 59-year-olds in two eastern provinces in Finland (Heistaro et al. 1998). The prevalence of chronic low back syndrome, diagnosed by a medical doctor, decreased during the 20-year period 1978/1980–2000 from 18.0% to 10.8% in men and from 16.8% to 11.0% in Finnish women aged 30 years or older (Kaila-Kangas 2007).

Contradictory results have been reported in the UK. The General Household Survey has shown fairly constant prevalence rates of self-reported chronic low back pain over a 9-year period (Symmons et al. 2002 as cited by McBeth and Jones 2007). Palmer et al. (2000) carried out two prevalence surveys at an interval of 10 years (1987/8–1997/8) among 20- to 59-year-olds geographically dispersed in Britain. Macfarlane et al. (2000) conducted two studies at a 7-year interval (1991–1998) with 18- to 80-year-old subjects from northwest England. The former study reported that the 1-year prevalence of back pain rose from 36.4% to 49.1%, whereas the latter reported a slight decrease of 1-month prevalence (from 26.1% to 22.6%).

The prevalence rates of radiographically detectable degenerative changes of the lumbar spine in a community sample over 34 years of age were described by Lawrence in 1969. The prevalence of disc degeneration (grades 1–4) was 51% in men aged 35–44, increasing to 91% at 65 or older. In women, the prevalence rates were 40% and 78%, respectively. The corresponding rates for more severe disc degeneration (grades 3 and 4) were 5% and 38% in men and 3% and 24% in women. One possible explanation for the gender difference is men's higher exposure to physical load; the prevalence was highest in men in physically strenuous occupations. A more sensitive method of detection, magnetic resonance imaging (MRI), has revealed high prevalence of disc degeneration even in symptom-free subjects, 6% in women 20 years of age or younger and 79% in those 60 years or older (Powell et al. 1986). In a large MRI study with 1,043 southern Chinese volunteers 18–55 years of age, 43% of the 18- to 29-year-olds had signs of lumbar disc degeneration. The proportion increased linearly with increasing age reaching 87% in 50- to 55-year-olds (Cheung et al. 2009).

The relationship between degenerative changes in the lumbar spine and low back pain is controversial. Van Tulder et al. (1997) found in their systematic review that radiographically detected degeneration of the lumbar spine was associated with non-specific LBP with odds ratios (OR) ranging from 1.2 to 3.3. No association was found between spinal deformities (spondylolysis and spondylolisthesis, spina bifida, transitional vertebra, etc.) and low back pain. In the MRI study by Cheung et al. (2009), self-reported lifetime history of LBP of more than 2 weeks duration and requiring physician consultation or treatment was associated with overall lumbar disc degeneration score (*OR*: 2.2, 95% confidence interval (CI): 1.4–3.4).

Risk Factors It has been estimated that worldwide 37% of LBP can be attributed to occupation. Work-related LBP was estimated to cause 818,000 disability-adjusted life years lost annually (Punnett et al. 2005). Occupational risk factors are to a great extent amendable and, accordingly, much research has been directed to identifying risk factors related to work.

Several comprehensive reviews on risk factors of back disorders have been published. In Table 54.2, the evidence for physical load is given based on the two most recent systematic reviews by Hoogendoorn et al. (1999) and Bakker et al. (2009). The former covered cohort and case-referent studies from 1949-1966 to 1997 and the latter prospective cohort studies of incident LBP (subjects free from LBP at baseline) from 1997 to 2007. These two reviews complement each other. Summarizing the results, the evidence is strong for heavy physical work (in terms of manual materials handling, lifting, forceful movements), bending and twisting postures at work, and whole body vibration being risk factors; evidence is strong for standing or walking and sitting at work not being risk factors; and the evidence is moderate for patient handling or nursing as risk factors.

Current knowledge on the mechanical risk factors of LBP is still mainly qualitative, and little is known about quantitative exposure-response relationships. The available evidence suggests that reduction of occupational physical load carries potential for the prevention of LBP, but data are sparse to define safety levels for exposure. Also little evidence exists of the effectiveness of workplace interventions affecting physical workload (Van der Beek et al. 2009).

Much research has addressed the role of work-related psychosocial factors in back pain, and several systematic reviews have been published. A EULAR Task Force (Macfarlane et al. 2009) evaluated the evidence of associations between back pain and workplace psychosocial factors based on published reviews (Table 54.3). The conclusions of the reviews were inconsistent. The reasons for this were differences in the dates of publication and the body of evidence; differences in methods used for synthesizing the evidence; differences in quality assessment; inclusion of studies (cross-sectional, case-control, cohort studies); and interaction between risk factors. The conclusions were most consistent for an association of

Table 54.2 Physical load as risk factor of back pain. Level of evidence according to two systematic reviews

	Hoogendoorn et al. 1999[a]	Bakker et al. 2009[b]
At work		
Manual materials handling/lifting/forceful movements	Strong evidence	Conflicting evidence
Bending and twisting/awkward posture	Strong evidence	Conflicting evidence
Heavy physical work	Moderate evidence	
Standing or walking	No evidence	Strong evidence of no association
Sitting	No evidence	Strong evidence on no association
Whole body vibration	Strong evidence	Conflicting evidence
Patient handling/nursing	Moderate evidence	Conflicting evidence
Professional sports		No evidence
Leisure time		
Sports activities/exercise	No evidence	Strong evidence of no association
Physical activity	No evidence	Conflicting evidence
Car driving	No evidence	
Sleeping		No evidence

[a]Cohort and case-referent studies from (1949-)1966–1997; three levels of evidence: strong, moderate, no ($\leq$1 study available or inconsistent findings in multiple studies)
[b]Prospective cohort studies on incident low back pain from 1997 to 2007; five levels of evidence: strong, moderate, limited (1 study available), conflicting (inconsistent findings), no (no studies found)

Table 54.3 Psychosocial work-related factors and back pain. Evidence of relationships extracted from seven reviews (Adapted from Macfarlane et al. (2009))

	Work demand		Low work	Low work	Low job
Review	High#	Low	control	support	satisfaction
Bongers et al. (1993)	+	+	0	+	0
Ferguson and Marras (1997)	+	0	+	+	+
Lagerstrom et al. (1998)	0	0	–	0	0
Hoogendoorn et al. (2000b)	0	0	0	+	+
Davis and Heaney (2000)	+	0	0	0	+
Linton (2001)	+	+	0	+	+
Hartvigsen et al. (2004)	–	–	–	0’	–

includes stress; + strong or moderate evidence; 0’ evidence against an association;
0 insufficient/no evidence; – not considered

back pain with high work demand, low job satisfaction, and low work support. The most recent review by Hartvigsen et al. (2004) was based on prospective cohort studies only and used defined methods for assessing the strength of association and consistency of evidence. This review concluded that there is moderate evidence of no association of low back pain with perception of work (6 variables, job satisfaction included), work organizational aspects (17 variables, work demand and work control included), and social support at work (8 variables). The evidence was insufficient for stress at work (5 variables).

For leisure time activities, the aforementioned reviews concluded that the evidence is strong of no association with back pain, for overall physical activity the evidence is conflicting, and for car driving and sleeping the evidence is lacking. Also for the psychosocial factors in private life, the evidence is insufficient (Hoogendoorn et al. 2000b).

Of person-related factors, physical fitness is often considered to have potential to protect from back disorders. Hamberg-van Reenen et al. (2007) carried out a systematic review of the evidence of low physical capacity as a predictor of future LBP. The evidence was summarized from prospective cohort studies. They found strong evidence of no relationship between trunk muscle endurance and the risk of LBP and inconclusive evidence for a relationship between trunk muscle strength or mobility of the lumbar spine and the risk of LBP. Recent meta-analyses by Shiri et al. (2010a, b) showed that overweight and obesity increase the risk of LBP with a dose-response relationship (pooled *OR*s ranging from 1.36 to 1.56 depending on the outcome measure) and that current and former smokers have a higher risk of LBP than never smokers (pooled *OR*s ranging from 1.30 to 2.14, accordingly). The association of current smoking with the incidence of LBP proved to be stronger among adolescents than adults and a dose-response relation was found among adolescents. The same group reported in a systematic review that overweight, long smoking history, high physical activity, and a high serum C-reactive protein level are associated with lumbar radicular pain or sciatica (Shiri et al. 2007). Another systematic review considered the association between atherosclerosis and disc degeneration and LBP (Kauppila 2009). It showed that smoking and high serum cholesterol level were quite consistently associated with disc degeneration and LBP.

Twin studies have indicated that both low back pain and disc degeneration have a substantial genetic component. Heritability estimates for low back problems have shown wide variation, explaining from 0% to 57% of the variance (Battié et al. 2007). One reason for the variation is in differences of the outcome definitions used in different studies. Pain is a very complex phenotype for genetic studies. Genetic effects can be linked to genes affecting intervertebral disc degeneration or immune response, or genes involved in pain perception, signaling, and psychological processing. Many candidate genes have been tested in association studies with varying results (Tegeder and Lötsch 2009). Much more research is needed to unravel the role of genetic and environmental factors in the manifestation of low back pain.

There is solid evidence of familial predisposition and a high heritable component in lumbar disc degeneration. The heritability estimates have ranged from 34% to 75% (Kalichmann and Hunter 2008a). A number of genes have been identified as being associated with intervertebral disc degeneration, including genes coding for collagen I, collagen IX (COL9A2, COL9A3), collagen XI (COL11A2), IL-1, aggrecan, vitamin D receptor, MMP-3, and CILP (Kalichmann and Hunter 2008b). Little is known of gene-gene or gene-environment interactions. More research is needed, including linkage studies and genome-wide assessment studies in different populations with full age range in order to understand the influence of genetic and environmental factors in this complex outcome.

Low back disorders are very common in population, and they are complex with multifactorial etiology. In spite of extensive research, primary causative mechanisms are not well understood, and a rather cynical view has been presented that due to the general nature and course of common LBP, there is limited scope for primary prevention (Burton et al. 2006). The body of evidence of the risk factors of low back disorders would warrant action to reduce harmful physical and psychosocial stressors. Also general health promotion in terms of physical exercise but avoiding traumatic injuries, quitting smoking, and controlling body weight may have a favorable effect on back morbidity. Evidence of the effectiveness of preventative interventions is sparse. The only intervention for which there is evidence of a positive effect in preventing low back pain is physical exercise. The effect size estimates range from 0.39 to 0.69 (Bell and Burnett 2009; Bigos et al. 2009). Negative trials have been reported of education, lumbar supports, shoe inserts, and reduced lifting programs. A recent review of the effectiveness of ergonomic workplace interventions on low back pain showed low to moderate quality evidence that physical and organizational ergonomic interventions are not more effective than no ergonomic intervention on LBP (Van der Beek et al. 2009).

54.2.2 Neck Disorders

In the ICD-10, neck disorders are classified as subcategories of back disorders or dorsopathies identifiable at the 3–4 digit level of the classification. The situation is analogous to that for back disorders: in majority of the cases a specific diagnosis cannot be set, but most of them fall into the category of non-specific neck pain (cervicalgia, “tension neck,” etc.). Accordingly, in most epidemiological studies, the health outcome has been unspecified neck pain.

Occurrence Neck pain is nearly equally common as back pain, but the impact in terms of short- and long-term disability is less. Studies based on the general population have shown large variation in prevalence estimates. This is partly accountable to variation in population sample characteristics, case definition, and case ascertainment. According to the review by Hogg-Johnson et al. (2008), lifetime prevalence estimates of any neck pain among adults ranged from 28% to 71%. Among adults 12-month prevalence of any neck pain ranged from 12% to 71% and among children from 34% to 71%. Most estimates of the 12-month prevalence were between 30% and 50%. One-month prevalence of any neck pain ranged from 15% to 45% among adults and from 4% to 8% among children and adolescents. When pain was further specified (frequent, of certain duration, interfering activities), the estimates were considerably lower. The review reported also incidence rate estimates from general population studies: estimates ranged from 0.055 per 1,000 person-years (disc herniation with radiculopathy) to 213 per 1,000 persons (self-reported neck pain).

The prevalence of neck pain among adults increases with increasing age to a peak in the middle age and declines thereafter. Among children and adolescents, the relationship between age and neck pain prevalence varies. Women tend to have higher prevalence of neck pain than men (Hogg-Johnson et al. 2008).

Data on the time trends of the occurrence of neck pain are scarce. In the Mini-Finland Health Survey (1978/80), age-adjusted 1-month prevalence of neck pain was 26.8% among men and 34.7% among women as compared with the Health 2000 Survey (2000–2001) where the corresponding figures were 25.8% and 39.9% indicating only a minor change during the 20 years in the population 30 years or older of age (Riihimäki et al. 2004). In these surveys, also the prevalence of physician-diagnosed chronic neck syndrome was estimated. Among men the prevalence decreased from 9.8% to 5.5% and among women from 13.9% to 7.3% showing a favorable development during the 20 years (Kaila-Kangas 2007).

Risk Factors Two recent systematic reviews addressed risk factors of neck pain, one in the general population (Hogg-Johnson et al. 2008) and the other one in workers (Côté et al. 2008). Both reviews covered literature from 1980 to 2006. A sufficient number of cohort studies were available to conduct the best evidence synthesis of risk factors of incident neck pain. The results are presented in Table 54.4.

For the general population (17 studies), evidence was found only for age, some psychological factors and exposure to passive smoking in childhood as being risk factors of neck pain. A number of other factors were considered in the studies, but the evidence was either preliminary or inconsistent. In worker populations (19 cohort studies, one RCT), a broader range of risk factors has been considered. Preponderance of evidence was found for age, history of musculoskeletal pains, high job strain, low social support at work, prolonged sedentary position, and repetitive or precision work; and evidence for low physical capacity, job insecurity, prolonged neck flexion, upper extremity posture, and key/mouse position.

Heritability of neck pain was estimated in a study of adult female twins in the UK (MacGregor et al. 2004). The estimates ranged from 35% to 58% depending on the definition of neck pain. In general, more severe neck pain definitions were associated with higher heritability estimates. Severe neck pain (lifetime occurrence of neck pain lasting more than 1 month and associated with disability) was most strongly associated with psychological distress (General Health Questionnaire, GHQ score $\geq$ 6%), but GHQ score accounted for only 1% of the genetic variation of severe neck pain. The association between severe neck pain and GHQ score was explained mostly by shared genetic factors than environmental factors. Severe neck pain was only weakly associated with degeneration of the cervical spine. Another twin study reported that genetic factors do not play an important role in the susceptibility to neck pain in persons 70 years of age or older (Hartvigsen et al. 2005).

Back and neck pain seem to share similar complex multifactorial etiology, in which both physical load and psychosocial factors have a role. Available evidence of risk factors suggests that workplace interventions aiming at reducing harmful physical load on the neck region and improving psychosocial aspects of work as well as physical capacity carry potential for the prevention of neck disorders.

Table 54.4 Evidence of risk factors of incident neck pain in general working populations. Data were extracted from the reviews based on cohort and case-referent studies and randomized controlled trials

Risk factor	General population[a] (Hogg-Johnson et al. 2008)	Worker population (Côté et al. 2008)
Demographic/socioeconomic factors		
Age	Evidence	Preponderance of evidence
Gender	Evidence varies	Preliminary evidence (F>M)
Marital status	Not associated	Evidence varies
Employment status	Not associated	
Number of children	Preliminary evidence	
Education		Evidence varies
Occupational class		Preliminary evidence
Duration of employment		Evidence varies
General health, comorbidity		
Poor self-assessed health	Preliminary evidence	Evidence varies
Obesity/BMI	Preliminary evidence (not a risk factor)	Evidence varies
Psychological factors	Evidence/preliminary evidence	
Low physical capacity		Evidence
History of musculoskeletal pains	Preliminary evidence	Preponderance of evidence
Health behaviour		
Smoking	Preliminary evidence	Preliminary evidence
Exposure to passive smoking in childhood	Evidence	
Leisure-time sport activities	Evidence varies	Evidence varies
Self-reported psychosocial/organizational factors at work		
High job strain		Preponderance of evidence
Low social support		Preponderance of evidence
Job satisfaction		Preliminary evidence (not a risk factor)
Job insecurity		Evidence
Stress at work		Preliminary evidence
Physical risk factors at work		
Prolonged sedentary position		Preponderance of evidence
Repetitive of precision work		Preponderance of evidence
Prolonged neck flexion		Evidence
Working with hands above shoulders		Evidence varies
Upper extremity posture		Evidence
Awkward postures		Preliminary evidence
Heavy physical work		Evidence varies
Keyboard/mouse position		Evidence
Intervention aimed at modifying workers' workstation and posture		Evidence for not reducing risk of neck pain

[a]The level of evidence was defined by this author (HR) based on the results presented by Hogg-Johnson et al. and applying the criteria of Côté et al.

The evidence of the effectiveness of preventive interventions is, however, scarce. In a recent review of ergonomic workplace interventions covering literature from 1988 to 2008, low-quality evidence, based on three RCTs, was found that physical ergonomic workplace intervention has no effect on short- or long-term incidence or prevalence of neck pain as compared with no such intervention (Van der Beek et al. 2009).

54.2.3 Osteoarthritis

Osteoarthritis (OA) is a degenerative process in the joints, in which loss of cartilage is seen as joint space narrowing on radiography. Bony changes include sclerosis in the subchondral bone, osteophyte formation, and bone cysts. Also soft tissue structures are affected. Clinical signs of osteoarthritis include pain and restricted range of motion in the joint, deformities, and sometimes instability. It is not clear whether osteoarthritis is a single disease or several disorders with a similar final common pathway (Felson 2000). Osteoarthritis can be classified as primary when the cause is unknown or secondary when the cause is known (e.g., joint injury). Generalized osteoarthritis may be a distinct disease where systemic genetic factors are more important than local mechanical factors. Support for such an entity was given by a study which showed that hand osteoarthritis predicts future hip or knee osteoarthritis (Dahaghin et al. 2005). In particular, osteoarthritis of the two large weight-bearing joints, the hip and knee, has a major clinical impact; it is common in elderly people causing pain and disability limiting daily activities and worsening the quality of life.

Occurrence The prevalence estimates of hip osteoarthritis (OA) vary considerably (Table 54.5) which is partly explained by different ascertainment methods of osteoarthritis and possibly also by different time periods. According to the Finnish population surveys, however, the overall prevalence of clinically defined hip osteoarthritis has remained quite stable from 1978/1980 to 2000 (Kaila-Kangas 2007): in men aged 30 years or older, the prevalence rates were 5.0% and 5.7% and in women 6.0% and 4.6%, respectively. The prevalence of hip OA increases steeply after the age of 55 years. In a study of the US population aged 45 years or older, the prevalence of radiographic hip osteoarthritis was 25.7% in men and 26.9% in women, whereas the prevalence of symptomatic osteoarthritis (symptoms and radiographic changes) was 8.7% in men and 9.3% in women (Lawrence et al. 2008).

The prevalence of knee OA also increases with age. It is distinctly more common in women than in men (Table 54.5). In various studies, knee OA was detected radiographically in 5–28% of men and in 8–40% of women aged 55–64 years. In the Framingham Study, repeated radiographs were taken in 1983–1985 (baseline) and in 1992–1993 (Felson et al. 1997). The mean age of the subjects without radiographically detectable knee OA at baseline was 70.5 years. During the follow-up, 11.1% of the men and 18.1% of the women developed moderate or severe knee OA.

Table 54.5 Prevalence of hip and knee osteoarthritis in general population studies (%)

	Finland Health 2000 Survey[a]		Northern England (Leigh)[b]		The Netherlands (Zoetermeer)[c]		USA HANES-I[d]	
Age	Men (n = 2,876)	Women (n = 3,480)	Men (n = 173)	Women (n = 207)	Men (n = 1,359)	Women (n = 1,598)	Both sexes (n = 2,358)	
Hip								
30–44	0.5	0.4						
45–54	1.8	0.7			2.5	2.3		
55–64	5.2	3.1	18.6	12.0	7.8	3.1	2.3	
65–74	12.2	11.5			8.5	12.5	3.9	
75–	24.1	21.4			10.5	19.1		
Total (age-adjusted)	5.7	4.6						
			(n = 550)	(n = 566)			Men (n = 2,498)	Women (n = 2,765)
Knee								
30/35–44	0.3	1.6	5.5	4.0			1.2	1.2
45–54	2.7	9.7	8.2	13.1	9.3	14.2	2.2	3.6
55–64	9.1	24.2	28.1	40.0	16.8	18.6	5.1	7.5
65–74	10.8	33.0	26.4	49.1	20.9	36.1	9.0	20.3
75–	21.0	38.4			22.1	45.9		
Total (age-adjusted)	6.1	8.0						

[a] Clinically defined (Kaila-Kangas 2007)
[b] Radiographically defined (Lawrence et al. 1966)
[c] Radiographically defined (Van Saase et al. 1989; right hip and knee)
[d] Clinically defined (Anderson and Felson 1988)

In Finland, the overall prevalence of men has been stable from 1980 to 2000 (6.1% and 6.1%), but among women, the prevalence has decreased from 15% to 8% (Kaila-Kangas 2007). In the studies of the US population aged 45 years or older, the prevalence of radiographic knee osteoarthritis has ranged from 18.6% to 24.3% in men and from 19.3% to 30.1% in women. The corresponding prevalence of symptomatic osteoarthritis (symptoms and radiographic changes) ranged from 5.9% to 13.5% in men and from 7.2% to 18.7% in women (Lawrence et al. 2008).

The prevalence of OA of the hand joints increases with increasing age, and it is more common in women than in men (Table 54.6). It occurs most frequently in the distal interphalangeal joints. In the Framingham Study, the prevalence of symptomatic hand OA was 26.2% in women and 13.4% in men aged at least 71 years (Zhang et al. 2002). The NHANES III study showed that in adults aged 60 years or more, the prevalence of radiographic and symptomatic hand osteoarthritis was 37.4% and 12.1%, respectively (Dillon et al. 2007).

Risk Factors Severe injury may be sufficient cause for osteoarthritis, but often there is interplay between systemic and local factors (Felson 2000). According to the prevailing understanding, mechanical load is important in the inception and development of OA, but systemic factors, such as obesity, bone density, hormonal status and estrogen use, nutritional factors, and genetics influence the process (Sowers 2001; Felson 2000). Extrinsic local factors include physical activity and injury, and the intrinsic factors include alignment (varus, valgus), muscle strength (quadriceps), joint laxity, and proprioception. Genetics seem to have a strong influence on the development of osteoarthritis (Felson 2000; Sharma 2001).

Several systematic reviews have synthesized the epidemiological evidence of risk factors of (incident) OA. Bierma-Zeinstra and Koes (2007) summarized the evidence for hip and knee osteoarthritis from systematic reviews. They found moderate to strong evidence that (high) physical workload is a risk factor of hip and knee OA, that high intensity sport activity is a the risk factor of hip OA, and that being overweight is a risk factor of clinical hip OA. Surprisingly, evidence for overweight/obesity as a risk factor for knee OA was not found, although traditionally the influence on the knee has been considered to be stronger than on the hip. In a more recent systematic review, three prospective cohort studies of the association between overweight/obesity and severe hip osteoarthritis (arthroplasty) were included in meta-analysis (Guh et al. 2009). The pooled incidence rate ratio (*IRR*) for overweight was 2.8 in men and 1.8 in women and for obesity 4.2 and 2.0, respectively. In one of the studies (Järvholm et al. 2005), the relative risk of severe hip osteoarthritis, as compared with normal weight, was 1.5 for overweight and 2.0 for obesity. For knee OA, the figures were 2.4 and 4.8, respectively. Those subject whose BMI was below normal weight range ($<20\,\mathrm{kg/m^2}$) hade a smaller risk of hip and knee OA than normal weight subjects. For hand OA, a systematic review found in its best evidence synthesis only moderate evidence of an association with overweight (Yusuf et al. 2010).

Another systematic review by Tanamas et al. (2009) showed that there is a paucity of studies on the association between knee malalignment and incident

Table 54.6 Prevalence of hand osteoarthritis (Kellgren grade ≥ 2) by age in two populations

	Men					Women				
Joint	30–44	45–54	55–64	65–74	75 +	30–44	45–54	55–64	65–74	75 +
Distal interphalangeal										
Finland[a]	0.4–2.3	3.7–10.9	11.4–26.7	21.3–36.4	24.0–51.0	0.2–1.2	5.7–11.8	24.2–40.0	40.3–62.1	48.2–68.9
Zoetermeer[b]	4.2	18.7	44.3	54.3	59.3	4.6	30.5	61.5	75.5	73.1
Proximal interphalangeal										
Finland	0–0.6	0–1.3	2.3–6.5	5.1–11.1	12.5–22.9	0–0.4	0.8–2.1	5.5–9.5	15.1–27.1	21.8–36.3
Zeotermeer	0.8	4.1	12.1	18.9	27.9	0.9	6.3	24.6	32.7	45.9
Metacarpophalangeal										
Finland	0–1.5	0–3.5	0.6–11.7	1.6–21.0	1.0–30.2	0–0.4	0.2–2.2	0–6.1	3.2–15.1	2.6–20.7
Zoetermeer	6.1	12.9	34.2	44.8	43.0	5.2	18.5	36.6	55.4	60.3
Carpometacarpal										
Finland	0.4	1.3–1.6	4.3–7.7	7.5–9.9	17.7–18.6	0.2	3.1–3.7	8.7–12.5	20.1–24.5	27.5–31.6
Zoetermeer	1.9	7.7	17.8	20.4	37.2	2.6	13.3	29.0	43.9	54.6

[a] Haara et al. (2003) Range of joint-specific prevalence rates. Finnish general population

[b] Van Saase et al. (1989) Overall prevalence rate. General population of Zoetermeer, The Netherlands

knee OA, and furthermore, the assumed relationship between the female hormonal aspects and OA of the hip, knee, and hand was not clearly observed in a systematic review by de Klerk et al. (2009).

Recreational physical activity is often recommended for the prevention of OA. In the Framingham Offspring Cohort, the effect of recreational physical activity on the development of knee osteoarthritis was studied (Felson et al. 2007): 1,280 subjects with mean age 53.2 years (range 26–81) were followed up for approximately 9 years. In this study, neither recreational walking, jogging, frequent working up a sweat, nor high activity levels relative to peers were associated with a decrease or increase in the risk of knee OA, even if the subjects were overweight. All earlier prospective cohort studies with inconsistent results have been smaller in size than the Framingham Study.

Hip injuries have been linked to hip OA in several cross-sectional and case-control studies, but prospective cohort studies are rare (Riihimäki and Viikari-Juntura 2000). Gelber et al. (2000) followed 1,321 former medical students, initially 22 years of age, for a median duration of 36 years. Hip and knee injury at cohort entry or during the follow-up increased the risk of OA in the corresponding joint.

In a large prospective cohort study of severe OA of the hip and knee among male employees in construction industry, great differences of risk were observed between occupational groups (Järvholm et al. 2008). In reference to white-collar workers, floor layers had the highest risk: 4.7-fold risk of knee and 1.6-fold risk of hip arthroplasty. Attributable fraction for work-related factors in severe knee osteoarthritis was estimated as 79% among floor layers.

Jensen gave a best evidence synthesis of the association between physical work demands and knee OA (Jensen 2008a) and hip OA (Jensen 2008b). Moderate evidence was found for a relationship between kneeling (*OR*s ranging from 1.1 to 2.9), and heavy lifting (*OR*s ranging from 1.9 to 2.5) and knee OA. For the combination of kneeling/squatting and heavy lifting, the evidence was also moderate and for climbing stairs limited. With regard to hip OA, the evidence was moderate to strong for heavy lifting. The burdens have to be at least 10–20 kg and the duration at least 10–20 years to give a clearly increased risk of hip OA. Working as a farmer for 10 years is associated with a doubled risk of hip OA (evidence is strong to moderate). Evidence for a relation between construction work and hip OA is limited, and there is insufficient or no evidence that climbing stairs or ladders causes hip OA.

Several studies have indicated that occupational factors such as repetitive work with hands increase also the risk of hand OA (Hadler et al. 1978; Lehto et al. 1990; Felson 2000). A study comparing female dentists and teachers showed that hand use may be protective of hand OA, whereas continuing joint overload may lead to joint impairment (Solovieva et al. 2005, 2006).

In classic twin studies, heritability of OA has varied between different sites, and it appears to be greater in women than men. The heritability estimates have been up to 60% for the hip and hand and 40% for the knee (Spector and MacGregor 2004; Felson 2009). Overall in the general population, little knee OA seems to be heritable, but in middle-aged women, heritability of bilateral knee OA reaches 40% (Felson 2009). A recent twin study reported that while genetic factors are important in the variation of the occurrence of OA at the hand, hip, and knee, no evidence was found of common or shared genetic factors determining the occurrence across these skeletal sites (MacGregor et al. 2009).

Several linkage and association studies have been carried out to identify susceptibility genes for OA. Linkage studies have implicated quantitative trait loci regions on chromosomes 2q, 9q, 11q, and 16p, among others, and association studies have implicated a great number of particular gene polymorphisms associated with an increased risk of OA (Spector and MacGregor 2004). Many reported associations by one group have, however, not been replicated by others. According to the review by Felson (2009), the most consistently reported genetic association is for FRZB (a gene coding secreted frizzle-related protein-3). This association has been shown especially for hip OA in women. The association of interleukin-1 (IL-1) gene family with OA has been found in several cohort studies, and genome-wide scans have also suggested that a gene associated with an increased risk of OA lies within the IL-1 cluster on chromosome 2q. For other genes, replications of the associations found for OA are not clear-cut.

It has become evident that multiple genes influence the development of OA, but the associations between single variants of genes and OA are modest. Valdes et al. (2008) used in their study on genetic risk of knee OA 36 single polymorphisms (SNPs) in 17 candidate genes, each of which had been found to be associated with OA in at least one study. Additive genetic risk equation for women included 12 SNPs and for men 8 SNPs, 4 of them being in the equations of both sexes. Women in the top quartile of the genetic risk variable had nearly ninefold risk of knee OA (OR: 8.98) as compared with women in the bottom quartile. The corresponding odds ratio was 5.06 for men. These results indicate that additive information from a number of genetic variants can predict a substantial proportion of risk of knee OA, but the results need to be confirmed in other study populations.

Based of the epidemiological evidence, weight control, avoidance of excessive physical load, and prevention of joint injuries should be the targets in preventive programs of OA. Moderate-level physical activity and exercise is often recommended for prevention of OA, but evidence for its effectiveness is lacking.

54.2.4 Upper Limb Disorders

Upper limb pain is common, and soft tissue syndromes rank high among workers' compensation claims. Non-traumatic upper limb disorders are also one of the leading causes of sick leaves. Over the years, many "umbrella" concepts have been used for these disorders, such as occupational overuse syndrome, repetitive strain injury, and cumulative trauma disorder, implying work-related etiology. Several specific disorders can be diagnosed: rotator cuff syndrome, epicondylitis, neural impingement syndromes (cubital tunnel syndrome, radial tunnel syndrome, carpal tunnel syndrome, Guyon canal syndrome), and tenosynovitis or peritendinitis of the wrist-forearm region (including de Quervain's disease). Yet a great proportion of the problem manifests only as non-specific symptoms. In this chapter, only the most common syndromes are discussed.

Occurrence Prevalence rates of the most common upper limb soft tissue syndromes obtained in general population and working population studies with clinical ascertainment are presented in Table 54.7. Shoulder syndromes have the highest prevalence varying from 2.1% to 8.3% among men and from 1.9% to 10.1% among women in the general population, depending on the specific syndrome studied. Prevalence of other upper limb syndromes ranged from 0.5% to 2.8% among men and from 1.1% to 9.2% among women in the general population. In working population studies, the prevalence rates were quite similar to those found in the general population.

Some data are available on the incidence of clinically confirmed upper limb syndromes. In a French working population survey (age 20–59 years), the Pays de la Loire Study, incidence of carpal tunnel syndrome was 1.00 per 1,000 person-years (0.6 in men and 1.4 in women). The incidence increased with increasing age in both genders (Ha et al. 2009). Nordstrom et al. (1998) reported a higher incidence rate, 3.5 per 1,000 person-years, in a US general population study. In this study, the case definition was based on medical records. In the study by Silverstein et al. (2006) on manufacturing and health care workers, 1-year cumulative incidence of rotator cuff tendinitis was 5.5% in the right and 2.9% in the left shoulder. For hand/wrist tenosynovitis, Kurppa et al. (1991) reported incidence rates ranging from less than 1–25 per 100 workers per year depending on gender and hand strain at work.

Criteria for the upper limb disorders are not standardized, and varying criteria have been used in different studies, which contributes partly to the obtained differences in the occurrence estimates.

Risk Factors All the upper limb disorders discussed above have been associated with repetitive movements of the hand, use of hand force, non-neutral wrist postures, and hand-arm vibration. According to previous reviews, the evidence of repetition and combination of these risk factors is strong for carpal tunnel syndrome but limited for the other risk factors. For epicondylitis and wrist-hand tendon disorders, the evidence is evaluated as moderate to limited for all these risk factors and their combinations (Vingård 2001a, b; Bernard 1997). In the review by the National Research Council, USA (2001), the AF estimates ranged from 44% (vibration) to 93% (repetition+force). The role of psychosocial factors was controversial. Carpal tunnel syndrome has been related to hormonal factors in older women, obesity and previous history of other musculoskeletal complaints.

In a 20-year follow-up of a Finnish general population cohort, self-reported exposure to repetitive movements and vibration at baseline increased the risk of chronic shoulder syndrome with *OR*s of 2.3 and 2.5, respectively (Miranda et al. 2008). The adverse effect of physical work exposures was detectable even among those older than 70 years at follow-up. Age and body mass index modified the effect of physical exposures.

Table 54.7 Prevalence (%) of upper limb soft tissue syndromes in general population and working population studies

	Shoulder syndrome		Lateral epicondylitis		Carpal tunnel syndrome (CTS)		Hand/wrist tenosynovitis	
	Men	Women	Men	Women	Men	Women	Men	Women
General population								
Kaila-Kangas (2007), Finland (age ≥ 30 years) (right/left)	5.8/3.7[a]	5.1/2.9[a]	0.6/0.5	0.8/0.5	1.2[e]/1.4[e]	3.5[e]/3.5[e]		
Walker-Bone et al. (2004), England (25–64 years)	4.5[b];8.2[c]	6.1[b];10.1[c]	1.3	1.1			0.5[g];1.1[h]	1.3[g];2.2[h]
Miranda et al. (2005), Finland (age 30–64 years)	2.1[b]	1.9[b]						
Shiri et al. (2006), Finland (age 30–64 years)			1.2	1.4				
Shiri et al. (2009), Finland (age ≥ 30 years)					2.1[e]	5.3[e]		
Atroshi et al. (1999), Sweden (age 25–74 years)					2.8[e];2.1[f]	4.6[e];3.0[f]		
de Krom et al. (1992), The Netherlands (age 25–74 years)					0.6[f]	9.2[f]		
Working population								
Roquelaure et al. (2006), France[i]	6.8[d]	9.0[d]	2.2	2.7	2.3[e]	2.7[e]	1.5[g,h]	2.6[g,h]
Silverstein et al. (2006), USA (right/left, M+F)[j]	7.6/5.0[b]							
Fan et al. (2009), USA (M+F)[j]			7.5					

[a] Chronic shoulder syndrome
[b] Tendinitis/rotator cuff tendinitis
[c] Adhesive capsulitis
[d] Rotator cuff syndrome
[e] Clinically diagnosed CTS
[f] Electrodiagnostically confirmed CTS
[g] de Quervain's disease
[h] Other tenosynovitis
[i] Mix of occupations
[j] Manufacturing, health care

In the Pays de la Loire Study, 11.3% of men and 15.1% of women were diagnosed to have an upper limb syndrome (rotator cuff syndrome, epicondylitis, cubital tunnel syndrome, extensor/flexor tendinitis/tenosynovitis, de Quervain's disease, or carpal tunnel syndrome). Manual workers had an excess risk of upper limb syndromes as compared with non-manual workers; the prevalence ratios ranged from 1.40 to 2.10 (adjusted for individual factors: age, obesity, diabetes, thyroid disease, osteoarthritis). Physical work exposures accounted for over 50% of the overall (any syndrome) differences between manual and non-manual workers, for 62% (in men) to 67% (in women) of rotator cuff syndrome and for 96% (in women) of carpal tunnel syndrome. The authors estimated that under low levels of physical exposures, up to 31% of cases among manual workers could have been prevented (Melchior et al. 2006).

Another recent cross-sectional community-based study has shown that using keyboard at least 1 h per day is associated with an increased risk of hand-wrist tendinitis (Walker-Bone et al. 2006) and the study by Silverstein et al. (2008) on manufacturing and health care workers that long duration of upper arm flexion and either forceful exertions ($OR = 2.4$) or forceful pinch ($OR = 2.7$) are risk factors of rotator cuff syndrome. In the latter study, high job structural constraints were significantly associated with an increased risk of rotator cuff syndrome, but for other psychosocial factors, the associations were marginal.

A substantial gradient across economic sectors and occupations exists for upper limb syndromes and also for workers compensation claims. According to the body of evidence, a large proportion of this morbidity could be prevented by reducing exposure to physical work factors. Evidence for the efficiency of interventions to prevent upper limb disorders is sparse, however.

In a systematic review by Amick et al. (2008), evidence synthesis was done of interventions aiming at primary (subjects without symptoms at baseline) and/or secondary prevention (subjects with symptoms at baseline) of neck and upper limb disorders among workers. Strong evidence of no effect was found for workstation adjustment alone, moderate evidence of no effect for biofeedback training and for job stress management training, and moderate evidence for a positive effect for arm supports. For other interventions (exercise, ergonomics training, workstation adjustment, alternative keyboards or pointing devices, cognitive behavioral training, rest breaks, participatory ergonomics, injury prevention programs miscellaneous work redesign), the evidence was limited, mixed, or insufficient. The number of high-quality studies is small (14 in the review by Amick et al.), and most of them have been carried out in office sector.

54.2.5 Osteoporosis

Osteoporosis has been defined as a systemic skeletal disorder characterized by low bone mass and microarchitectural deterioration of bone tissue, with a consequent increase in bone fragility and susceptibility to fracture (Consensus Development Conference 1993). Osteoporosis is often identified by the occurrence of characteristic low trauma fractures (fragility fractures), the most serious of which are

Table 54.8 Prevalence of osteoporosis among an age-stratified sample of residents of Rochester, Minnesota. World Health Organization criteria (2.5 SD or more below sex-specific young normal mean) (Adapted from Melton (2003))

Age (years)	*n*	Hip (%)	Lumbar Spine (%)	Wrist (%)	Hip or spine or wrist (%)
Postmenopausal women					
50–59	50	4.0	2.0	6.0	8.0
60–69	50	10.0	8.0	28.0	30.0
70–79	51	19.6	17.6	56.9	56.9
>80	50	40.0	4.0	78.0	82.0
Total	201	13.6[a]	7.7[a]	32.9[a]	34.7[a]
Men					
50–59	49	12.2	2.0	4.1	12.2
60–69	50	16.0	0	4.0	18.0
70–79	51	15.7	2.0	11.8	21.6
>80	50	26.0	2.0	30.0	40.0
Total	200	15.8[a]	1.4[a]	8.8[a]	19.4[a]

[a] Prevalence per 100, age-adjusted to the 1990 US white population ≥ 50 years old

fractures of the hip and vertebrae. One of the strongest risk factors of fractures is low bone mineral density (BMD). High-precision methods to identify BMD are dual energy X-ray absorptiometry (DXA) and quantitative computed tomography. DXA provides an accurate, reproducible, and safe means of measuring BMD, and it is widely accepted as the reference method (Pafumi et al. 2002). For screening purposes, quantitative ultrasonometry (QUS) of the calcaneal bone can be used.

The World Health Organisation (1994) has recommended four categories to classify osteoporosis: normal, BMD not more than 1 standard deviation (SD) below the mean of young adults; and osteopenia, BMD between 1 and 2.5 SD below the mean of young adults; osteoporosis, BMD more than 2.5 SD below the mean of young adults; severe or established osteoporosis, BMD more than 2.5 SD below the mean of young adults and one or more fragility fractures. The criteria are useful in population-based surveys but less useful in assessing individual risks (Jordan and Cooper 2002).

Occurrence In Table 54.8, the prevalence of osteoporosis is presented by age for men and women 50 years or older in a sample of the general population, Rochester, Minnesota. A clear increase with increasing age is seen, but it is less steep for the lumbar spine than for other sites. This may be due to age-related artifacts that mask bone loss from the vertebral body (Melton 2003). Prevalence of osteoporosis in the lumbar spine, wrist, and all sites combined was higher in women than in men, but the genders had similar prevalence rates for the hip. In some other studies, in which the same absolute bone density cut-off level for men and women has been used, lower prevalence rates have been obtained for men (Melton 2003).

Incidence rates of fractures were estimated for the general UK population during the 10-year period 1988–1998 (Van Staa et al. 2001). For women aged 50 years or older, the incidence rate of hip fractures was 37.2 per 10,000 person-years, and the incidence of distal forearm fractures was 54.4 per 10,000 person-years. For men, the corresponding figures were markedly lower, 11.1 and 11.2 per 10,000 person-years. For 50-year-old women, the estimated remaining lifetime risk of a hip fracture was 11%, of the distal forearm 17%, and of the spine 3%. The corresponding estimates for men were 3%, 3%, and 1%. For the spine, the figures may be underestimated because only clinically verified fractures were recorded. In the European Vertebral Osteoporosis Study, overall 12% of women and men aged 50–79 years were reported to have radiographic evidence of vertebral deformities indicating vertebral osteoporosis (O'Neill et al. 1996).

In the Health 2000 Survey (a representative sample of the Finnish general population 30 years or older), QUS at the heel (calcaneal bone) was used to detect osteoporosis (Kaila-Kangas 2007). The mean of those with good or rather good perceived health in the age group 30–44 years of the same gender was used as the reference for women and men. The estimated average BMD was roughly the same in women and men aged 30–54 years (about 0.56–0.57 g/cm^2). In the older age groups, BMD decreased with increasing age and the change was more pronounced among women than men, as expected. In women 80 years or older, the mean BMD was 0.32 g/cm^2 and in men 0.46 g/cm^2. A remarkable difference between genders was detected, when reference value mean-2SDs was used: in the age group 75–84 years, the prevalence of those with a lower value was 36.8% in women and 6.2% in men, and in the age group 85+, the figures were 62.8% and 14.7%, respectively.

Risk Factors Espallargues et al. (2001) carried out a systematic review on bone-mass-related risk factors for fracture. From 94 cohort and 72 case-control studies, they identified 80 risk factors. For more than half of the risk factors, scientific evidence was insufficient or contradictory. Aging (>70–80 years), low body weight, weight loss, physical inactivity, the consumption of corticosteroids or anticonvulsants, primary hyperparathyroidism, diabetes mellitus type I, anorexia nervosa, gastrectomy, pernicious anemia, and previous osteoporotic fracture were associated with high risk (relative risk (RR) ≥ 2) of fractures. Moderate risk ($1 < RR < 2$) was found for female gender, active smoking, low sunlight exposure, family history of osteoporotic fracture, surgical menopause, early menopause, late menarche, short fertile period, no lactation, low calcium intake (<500–850 mg/day), hyperparathyroidism, hyperthyroidism, diabetes mellitus type II, and rheumatoid arthritis.

Early developmental factors such as tall maternal height and low rate of childhood growth seem to have an influence on later osteoporosis risk (Cooper et al. 2001). Sufficient calcium intake and physical activity during the growing years of young people are major determinants of bone mass which reaches its peak around the age of 25. A meta-analysis of the effect of calcium intake on bone mass in 18- to 50-year-olds showed that calcium intake of about 1,000 mg/day can prevent the bone loss of 1% of bone per year at all bone sites except in the ulna in premenopausal women. For men, too few studies were available to draw conclusions (Welten et al. 1995). Another more recent meta-analysis studied the use of calcium or calcium in combination with vitamin D supplementation to prevent bone loss and fractures in people aged 50 years or older (Tang et al. 2007). Supportive evidence was obtained for the use of calcium or calcium in combination with vitamin D. The combination was associated with a reduced bone loss of 0.54% at the hip and 1.19% in the spine. The overall relative risk of fracture was 0.90 for calcium only and 0.87 for the combination.

Meta-analyses on the effect of physical exercise on bone loss in postmenopausal women have shown that regular walking has no significant effect on preservation of BMD at the spine, but it has a significant positive effect at the femoral neck (Martyn-St James and Carroll 2008), whereas mixed loading exercise programs combining jogging with other low-impact loading activities and programs mixing impact activities with high-magnitude resistance training appear to be effective both at the hip and spine (Martyn-St James and Carroll 2009). However, Moayyeri (2008) stated in his review that the positive effects of physical activity on BMD and bone quality are of a questionable magnitude for reduction of fracture risk. The favorable effect of moderate to vigorous physical activity on hip fracture risk reduction (45% among men and 38% among women) was suggested to be mediated by a reduced risk of falling among physically active people.

Osteoporosis is strongly influenced by genetic factors. Twin and family studies have shown that 50–85% of the variance in bone mass is genetically determined (Williams and Spector 2006). Most studies on genetics of osteoporosis have focused on the phenotype of BMD, and it has been demonstrated that heritability of BMD is polygenic (Duncan and Brown 2008). In most studies, a candidate gene approach has been used. The genes most often studied are vitamin receptor gene (VDR), the collagen type 1 alpha 1 gene (COLIA1), and estrogen receptor gene (ER) alpha. They have been found to influence BMD, but the effects are modest and probably account for less than 5% of the heritable proportion of BMD (Williams and Spector 2006). Progress in understanding the genetic influences in osteoporosis is expected from ongoing genome-wide association studies (Duncan and Brown 2008).

With the expected demographic shift toward older age groups in the industrialized countries, the number of affected people, women in particular, will increase substantially. Prevention of osteoporosis is an important public health issue. Optimizing maternal nutrition and fetal growth, improving calcium intake and general nutrition in growing children, and increasing the exercise level throughout the population have been suggested to be included in the preventive population strategies (Walker-Bone et al. 2002).

54.3 Methodological Problems in Epidemiological Research

54.3.1 Aspects of Study Design

Most epidemiological studies have used cross-sectional study design, but the number of cohort and case-control studies has been increasing as can be seen from recent systematic reviews. A wide range of potential determinants have been explored in cross-sectional settings, but advancing the knowledge of risk factors of musculoskeletal disorders requires more rigorous study designs.

Case-control studies of musculoskeletal disorders are rare because of the difficulties in case definition and catchment of cases (cf. chapter ▶Case-Control Studies of this handbook). The majority of people with musculoskeletal problems are treated in primary health care. Generally accepted diagnostic criteria are not available, and the use of diagnostic labels varies remarkably among practitioners. In some studies, cases have been identified from health care providers (e.g., Vingård et al. 2000) or from registers (e.g., Heliövaara et al. 1987). In such studies, one needs to be aware of potential bias in etiological considerations due to differential proneness across exposure groups to file a claim or seek care or to get treatment. This bias can be minimized by restricting the cases to only severe cases, which will be treated in a similar way irrespective of, e.g., social class or demands on physical performance (e.g., Manninen et al. 2002).

Another question to consider in case-control studies is the validity and relevance of exposure assessment (cf. chapter ▶Exposure Assessment of this handbook). Often the only feasible option for data collection is to use retrospective questionnaires or interviews providing data susceptible to recall error. This error can be differential between cases and referents. Due attention should be paid to the relevant time windows for exposure assessment in reference to the inception of the disease of interest and induction period. In very few studies, this matter has been considered, but some examples can be found (Vingård et al. 1991a). In cases where the inception and induction periods are not known, assumptions can be made and tested. For studying transient triggering factors of musculoskeletal disorders, a case-crossover design may be the approach of choice (cf. chapter ▶Modern Epidemiological Study Designs of this handbook). This approach is appropriate, for instance, for sudden-onset overexertion injuries. Even though degenerative process of the spine, chronic in nature, may be the underlying cause for back and neck disorders, these disorders characteristically manifest as painful episodes. Little is known of the possible triggering causes of such spells.

During the past 10 years, an increasing number of cohort studies on musculoskeletal disorders have been published (cf. chapter ▶Cohort Studies of this handbook). The exposures of interest and other determinants have been measured at baseline and used as predictors of new-onset musculoskeletal pain or incident cases of musculoskeletal disorders. According to the general principles of etiological cohort studies, persons already afflicted by the index disorder should be excluded. This simple principle turns out to be quite difficult to follow for disorders with a chronic underlying process, such as disc or cartilage degeneration, which manifest as intermittent pain episodes. The date of inception for such disorders is not easy to define, and operational definitions, preferably based on assumed or known pathomechanisms, should be employed. As an example, for osteoarthritis, a natural proxy is the time of the first occurrence of “arthritic” pain in the afflicted joint.

In prospective cohort studies, health outcome can be assessed using repeated surveys including symptom questionnaires or interviews, imaging or functional performance tests, or clinical tests. When available, register data are useful particularly in large-scale epidemiological studies (e.g., Heliövaara et al. 1987) (cf. chapter ▶Use of Health Registers of this handbook).

54.3.2 Health Outcome Assessment

A major difficulty in conducting systematic literature reviews and meta-analyses of the epidemiological evidence of musculoskeletal disorders is the lack of generally accepted diagnostic or classification criteria of health outcomes. Many of the available criteria are based on a consensus of a group of experts, and validity of the classification criteria has been investigated surprisingly little even for the most common disorders. Among others, this was the conclusion of Reijman et al. (2004) in their investigation of validity, reliability, and applicability of seven definitions of hip osteoarthritis used in epidemiological studies. Likewise, Schiphof et al. (2008) concluded, based on their systematic appraisal of 25 different classification criteria of knee osteoarthritis, that more research is needed to reach uniformity in the classification criteria to be used in epidemiological research. Their recommendation was that meanwhile separate scoring of clinical and radiographical findings, overall scoring and pain registration should be used. In a systematic review and meta-analysis of etiological risk factors for low back pain, Griffith et al. (2007) found altogether 132 outcome definitions in 55 studies. Using an administered Delphi process, a group of experts categorized the definitions and reached a consensus for 20 sets of outcomes (3 sets of pathology outcomes, 2 sets of functional limitation, 2 sets of participation in work, and 13 sets of symptom and care seeking outcomes). These findings emphasize the complexity of low back pain as an outcome in epidemiological studies but also the urgent need for uniform and valid outcome definitions.

Questionnaire Surveys The occurrence of unspecified pain has been the most common outcome measure in epidemiological studies on musculoskeletal disorders. The occurrence parameter has been 1-, 3-, or 12-month or lifetime period prevalence or cumulative incidence. One of the major drawbacks of such symptom data is that mild pain due to occasional overexertion is not differentiated from more chronic and severe conditions, although this would be important in both etiological and prognostic studies. Pain perception is subjective and as such person-dependent and, therefore, modified by several factors like prior pain experience, culture, coping mechanisms, etc. Same personal features can influence other measures based on subjective perception, psychosocial factors as an example. Pain experience may also depend on the level of physical activity at work and leisure, because physical activity provokes musculoskeletal pain. As a result, spurious associations between potential determinants and musculoskeletal disorders can be obtained.

No generally agreed symptom questionnaires are available. One of the most commonly employed forms is the Nordic questionnaire (Kuorinka et al. 1987). Lack of consensus in defining such fundamental concepts as new-onset cases, acute, recurrent, and chronic cases, and an episode hampers the development of

Table 54.9 Characteristics of back pain and their measurement (Adapted from Croft and Raspe (1995))

Characteristic		Measurements
Back pain	Location Intensity	Pain drawing Visual Analogue Scale Numerical rating Verbal rating
	Quality Temporality	McGill Short Form Duration of current episode Number of days of back pain in defined period (6 or 12 months)
Bodily symptoms	Other pains Other complaints	Pain drawing Area checklist MSPQ[a] Symptom-Checklist 90R
Psychological features	Emotions	General Health Questionnaire Zung Depression Index Hospital Anxiety Depression Scale
	Cognitions	Pain-related control scale Fear Avoidance Beliefs
	Behavior Psychiatric disorders	Inappropriate symptoms DSM-III[b]
Signs	Systemic Local back Local other	Fever, ESR, weight loss Tender points SLR[c]: active, passive to reproduce leg pain Femoral stretch Schober’s test Lateral flexion Inappropriate signs

[a] MSPQ = Modified Somatic Perceptions Questionnaire
[b] DSM III = Diagnostic and Statistical Manual of the American Psychiatric Association (American Psychiatric Association 1980)
[c] SLR = Straight leg raise

epidemiological research focusing on pain and other symptoms. For low back pain, proposals for the definition of episodes (de Vet et al. 2002), recurrence of episodes (Stanton et al. 2009), and back pain in prevalence studies (Dionne et al. 2008) have been made. Also a scoring system for pain intensity has been presented (von Korff et al. 1990). Nachemson (2000) proposed time limits to define acute (0–3 weeks’ duration of pain or disability), subacute (4–12 weeks’ duration of pain or disability), and chronic back or neck pain (more than 12 weeks’ duration of pain or disability) and recurrent problems (patients seeking help after at least 1 month of not seeking care or not being on sick leave after at least 1 month of working).

Table 54.9 summarizes the characteristics of pain and indicates the most commonly used measuring methods (for references, see the original article).

Clinical Tests Clinical examination, laboratory tests, and functional performance tests usually do not provide very useful information for the case definition of back and neck disorders. For the back, a comprehensive review of the reliability of different function measurements concluded that recommendations for tests to be used in epidemiological studies cannot be made (Essendrop et al. 2002). Another systematic review of physical examination of non-specific low back pain also demonstrated low reliability (May et al. 2006). For upper limb disorders, a criteria document has been published proposing diagnostic criteria based on symptoms and clinical findings (Sluiter et al. 2001). For osteoarthritis, clinical examination has been used for classification, e.g., the American College of Rheumatology has presented classification criteria (www.rheumatology.org/research/classification).

Imaging Radiological site-specific definition of osteoarthritis is recognized as the method of choice in epidemiological studies of osteoarthritis (Hart and Spector 1955, 2000). Exposure to radiation is a matter of concern especially when taking radiographs of the hip, pelvic, or lumbar spine area. Kellgren and Lawrence presented already in 1957 classification criteria (Kellgren and Lawrence 1957; Kellgren 1963), which have been widely used. In this system, osteoarthritis is classified using grades from 0 to 4. The grading is based on a presumption of sequential appearance of osteophytes, loss of joint space, subchondral sclerosis, and cyst formation, but it relies heavily on the presence of osteophytes. Criticism has been presented against this classification, and there is growing consensus that semiquantitative grading of osteophytes and joint space narrowing, using validated atlases, should be separately recorded. Based on these data, also aggregate measures can be constructed (Hart and Spector 1955, 2000).

For the imaging of the spine, magnetic resonance imaging (MRI) is a sensitive method to study the degenerative process of spinal tissues, the intervertebral disc, in particular. Compared to traditional radiography, early stages of degenerative process can be detected. No known adverse health effects are involved, but high costs and availability of the equipment limit the use of MRI in large-scale epidemiological studies. No generally agreed classification criteria exist, but for disc degeneration, a system based on the morphology of the disc (inhomogeneity of disc structure, clarity of the distinction between the annulus and nucleus, and collapse of the disc space) has been proposed (Pfirrmann et al. 2001). Instead of using aggregate grading systems combining different features of disc degeneration, it seems advisable to record the features (morphology, signal intensity, bulging) separately in epidemiological studies. For other musculoskeletal disorders, MRI has not been utilized in many studies, and classification criteria are not available.

Bone Mineral Density Measurements Bone mineral density (BMD) can be measured with high-precision using dual-energy X-ray absorptiometry (DXA). It is based on the attenuation of X-rays passing through the bone. With this technique, it is possible to detect bone mass reduction that carries increased risk of fracture. DXA is widely accepted as the reference method to diagnose osteoporosis and fracture risk (Pafumi et al. 2002). The use of DXA in large epidemiological studies is limited by the availability of the equipment that is usually located in specialized centers. Another method, quantitative ultrasonometry (QUS) of the calcaneal bone, has several advantages. The equipment is cheaper, it is transportable, and the execution of the measurement is rapid. However, QUS is more suitable to assess bone density in younger (up to 55 years) than old people (Pafumi et al. 2002).

Registers Register data of occupational injuries, users of social security benefits such as sick leave and disability pension, or of users of health care services (hospital discharge registers, registers from health care providers) can be used for descriptive epidemiology to find gradients across different population groups (e.g., Leino-Arjas et al. 2002; Vingård et al. 1991b; Silverstein et al. 2002). Register-based case assessment of musculoskeletal disorders in etiological studies carries a potential for bias due to differential proneness across population groups to file a workers' compensation claim, seek or get treatment, to go on sick leave or pension. Another weakness is non-uniform use of diagnostic labels among medical doctors. In some countries, a national register of arthroplasties is being kept, making it easy to identify severe cases of osteoarthritis (Manninen et al. 2002).

54.3.3 Exposure Assessment

Assessment of exposure to mechanical load is a challenge in epidemiological studies. For many other risk factors, such as anthropometric measures and lifestyle factors, standardized and validated methods are available. Also for psychosocial factors, validated questionnaires are available, as an example the widely used Job Content Questionnaire to assess psychosocial job characteristics (Karasek et al. 1998).

Exposure to mechanical load as a risk factor is conceptually problematic. Everybody is "exposed" to materials handling, awkward postures, repetitive movements and many also to vibration, the known risk factors of musculoskeletal disorders. Sedentary lifestyle is becoming more common and has raised concern of too low "exposure" levels to mechanical load leading to deterioration of people's physical condition. A crucial question is, at what level mechanical load starts having adverse effects on the musculoskeletal system.

Mechanical exposure involves exertion, motion, and posture. Each of these can be characterized by level (amplitude), frequency (repetitiveness), and duration. Dynamic movement characteristics, velocity and acceleration, may also be important. Furthermore, for different musculoskeletal disorders, the relative importance of peak and cumulative exposures varies. This makes mechanical exposure very complex to assess, and for the planning of the assessment strategy, one should have prior knowledge or hypotheses of the pathomechanisms of the disorders in question. Using biomechanical modeling, the internal exposure can be estimated. Also common metrics have been proposed for the assessment of physical workload (Wells et al. 1997).

In epidemiological studies, mechanical exposure can be assessed using self-reports, observation methods, or direct measurement. Winkel and Mathiassen (1994) have described the feasibility of different assessment methods. Capacity, versatility, and generality decrease, whereas exactness and cost increase from self-reports to observations to direct measurements. With modern sophisticated measuring technology, it is possible to get accurate measures for current exposure to various dimensions of physical load at work. Direct measurements may not be feasible for large studies, and the accuracy of measurements is of value only if the current exposure is relevant with regard to study objective. Also, if current exposure can be considered as a proxy of past exposure, direct measurements are warranted. However, often this may not be the case, and other means of past exposure assessment, such as interview or expert assessment based on occupational history, must be utilized. Several validation studies of self-assessment of physical exposures, have shown that the accuracy is not always very good, but for some exposures such data are usable in epidemiological studies (Mortimer et al. 1999; Viikari-Juntura et al. 1996).

Different techniques to assess posture, including observational methods, video-taping and computer-aided observational methods, direct measurements, as well as self-reports, have been reviewed and discussed by Li and Buckle (1999). Precision, cost, and feasibility of different techniques to assess postural load, mechanical load, psychophysical load, and physiological load in epidemiological studies have been compared by Burdorf et al. (1997). Also Van der Beek and Frings-Dresen (1998) have reviewed mechanical exposure assessment methods and strategies in epidemiological studies. Direct measurements, such as electromyography, can be used for more accurate assessment. Methods for data reduction have been developed, such as exposure variation analysis (EVA, Mathiassen and Winkel 1991) or its modification, clustered EVA (Anton et al. 2003).

The spectrum of different dimensions of mechanical load and measurement methods poses a problem in conducting systematic reviews and meta-analyses of the association between mechanical load and musculoskeletal disorders. Griffith et al. (2008) developed common metrics for the translation of literature-based workplace mechanical exposures for use in meta-analyses on low back pain among workers. They developed seven-point scales for trunk posture, weight lifted or force exerted and spinal loading, and estimated both peak and cumulative loads. Such metrics can help comparison across studies, but in the future, more consistent measures of mechanical exposure are needed.

54.4 Conclusions

In epidemiology of musculoskeletal disorders, steady progress has been made during the past decades. Even though cohort and case-control studies are still sparse, their number is increasing. Systematic reviews have collected the available evidence of the risk factors of musculoskeletal disorders. Major risk factors have been identified qualitatively, but little is known of the exposure-effect relationships.

More studies with rigorous study designs are needed with valid methods to assess both exposures and health outcomes.

The lack of knowledge on exposure-effect relationships makes it difficult to plan preventative programs and agree on standards and guidelines to reduce the risk of musculoskeletal disorders. The knowledge base is, however, sufficient to plan interventions aiming at prevention. In workplace interventions, a good starting point is to reduce exposure to highest loads. Reviews of intervention studies have revealed that high-quality studies are rare and all the reviews conclude that randomized controlled trials are needed to provide convincing evidence of the effectiveness of the preventative interventions.

For low back and neck pain, it has been shown that both physical load and psychosocial factors play a role. When subjective perception of non-specific pain is the health outcome, the question remains, how much physical load acts as a symptom provoking or aggravating factor and how much it is a causal factor for back morbidity. Likewise, it can be questioned to what extent psychosocial factors modify the pain perception of the subjects. This question is relevant because in experimental studies, it has been shown that pain perception of individuals varies very much when exposed to a standardized pain-inducing stimulus. For the future development of epidemiology of back and neck disorders, it is important that effort is put to identifying more specific and also clinically relevant entities of low back and neck disorders. For upper limb disorders, better diagnostic criteria are available, but the specific diagnoses cover only a part of all upper limb complaints. It is equally important to develop exposure assessment methods that are valid and feasible also for large-scale population-based studies.

Recent studies have revealed that the prevalence of low back and neck disorders starts to increase already in the adolescence. There also seems to be an association between the development of disc degeneration and low back pain among the young (Salminen et al. 1999). These results suggest that follow-up studies of young cohorts are needed to learn more about the temporal relationships between low back and neck pain and the degenerative process of the intervertebral disc. Another question of interest is the inception and natural course of low back and neck disorders from the adolescence through adulthood to older age.

Genetic epidemiology is emerging also in the field of musculoskeletal disorders. Twin studies have provided information of the familial and genetic influences on musculoskeletal disorders of the back and joints, in particular. Many candidate genes have been detected, but much remains to be done to identify the genes involved and their influences. A future challenge is to learn about the interactions between genetic susceptibility and environmental factors.

References

Amick BC, Kennedy CA, Dennerlein JT, Brewer S, Catli S, Williams R, Serra C, Gerr F, Irvin E, Mahood Q, Franzblau A, Van Eerd D, Evanoff B, Rempel D (2008) Systematic review of the role of occupational health and safety interventions in the prevention of upper extremity musculoskeletal symptoms, signs, disorders, injuries, claims and lost time. Institute for Work and Health, Toronto

Anderson JJ, Felson DT (1988) Factors associated with osteoarthritis of the knee in the first national health and nutrition examination survey (HANES I). Evidence for an association with overweight, race, and physical demands of work. Am J Epidemiol 128:179–189

Anton D, Cook TM, Rosencrance JC, Merlino LA (2003) Method for quantitatively assessing physical risk factors during variable noncyclic work. Scand J Work Environ Health 29: 354–362

Atroshi I, Gummesson C, Johnsson R, Ornstein E, Ranstam J, Rosén I (1999) Prevalence of carpal tunnel syndrome in a general population. JAMA 281:153–158

Bakker EW, Verhagen AP, van Trijffel E, Lucas C, Koes BW (2009) Spinal mechanical load as a risk factor for low back pain: a systematic review of prospective cohort studies. Spine 34: E281–E293

Battié MC, Videman T, Levalahti E, Gill K, Kaprio J (2007) Heritability of low back pain and the role of disc degeneration. Pain 131:272–280

Bell JA, Burnett A (2009) Exercise for the primary, secondary and tertiary prevention of low back pain in the workplace: a systematic review. J Occup Rehabil 19:8–24

Bernard B (ed) (1997) Musculoskeletal disorders and workplace factors. A critical review of epidemiologic evidence for work related musculoskeletal disorders of the neck, upper extremity, and low back pain. US Department of Health and Human Services CDC (NIOSH), Cincinnati. DHHS (NIOSH) Publication No. 97–141

Biering-Sorensen F (1982) Low back trouble in a general population of 30-, 40-, 50-, and 60-years-old men and women. Study design, representativeness and basic results. Dan Med Bull 29:289–299

Bierma-Zeinstra SM, Koes BW (2007) Risk factors and prognostic factors of hip and knee osteoarthritis. Nat Clin Pract Rheumatol 3:78–85

Bigos SJ, Holland J, Holland C, Webster JS, Battié M, Malmgren JA (2009) High-quality controlled trials on preventing episodes of back problems: systematic literature review in working-age adults. Spine 9:147–168

Blyth FM, Jones GT, Macfarlane GJ (2009) Musculoskeletal health – how early does it start? Rheumatology 48:1181–1182

Bongers PM, de Winter CR, Kompier MA, Hildebrandt VH (1993) Psychosocial factors at work and musculoskeletal disease. Scand J Work Environ Health 19:297–312

Burdorf A, Rossignol M, Fatallah FA, Snook SH, Herrick RF (1997) Challenges in assessing risk factors in epidemiologic studies on back disorders. Am J Ind Med 32:142–152

Burton K, Müller G, Balagué F, Cardon G, Eriksen HR, Hänninen O, Harvey E, Henrotin Y, Indahl A, Lahad A, Leclerc A, van der Beek A (2006) Chapter 2. European guidelines for prevention in low back pain: november 2004. Eur Spine J 15(suppl 2):S136–S168

Cheung KM, Karppinen J, Chan D, Ho DW, Song YQ, Sham P, Cheah KS, Leong JC, Luk KD (2009) Prevalence and pattern of lumbar magnetic resonance imaging changes in a population study of one thousand forty-three individuals. Spine 34:934–940

Consensus Development Conference (1993) Diagnosis, prophylaxis and treatment of osteoporosis. Am J Med 94:646–650

Cooper C, Eriksson JG, Forsen T, Osmond C, Tuomilehto J, Barker DJ (2001) Maternal height, childhood growth and risk of hip fractures in later life. Osteoporos Int 12:623–629

Côté P, van der Velde G, Cassidy JD, Carroll LJ, Hogg.Johnson S, Holm L, Carragee EJ, Haldeman S, Nordin M, Hurwitz EL, Guzman J, Peloso PM (2008) The burden and determinants of neck pain in workers. Results of the Bone and Joint Decade 2000–2010 Task Force on Neck Pain and Its Associated Disorders. Spine 33(4S):S60–S74

Croft P, Raspe H (1995) Back pain. Baillière's Clin Rheumatol 9:565–583

Dahaghin S, Bierma-Zeinstra SMA, Reijman M, Pols HAP, Hazes JMW, Koes BW (2005) Does hand osteoarthritis predict future hip or knee osteoarthritis? Arthritis Rheum 52:3520–3527

Davis KG, Heaney CA (2000) The relationship between psychosocial work characteristics and low back pain: underlying methodological issues. Clin Biomech 15:389–406 de Klerk BM, Schiphof D, Grieneveld FP, Koes BW, van Osch GJ, van Meurs JB, Bierma-Zeinstra SM (2009) No clear association between female hormonal aspects and osteoarthritis of the hand, hip and knee: a systematic review. Rheumatology 48:1160–1165

de Krom MCTFM, Knipschild PG, Kester ADM, Thijs CT, Boekkooi PF, Spaans F (1992) Carpal tunnel syndrome: prevalence in the general population. J Clin Epidemiol 45:373–376

de Vet HCW, Heymans MW, Dunn KM, Pope DP, van der Beek AJ, Macfarlane GJ, Bouter LM, Croft PR (2002) Episodes of low back pain. A proposal for uniform definitions to be used in research. Spine 27:2409–2416

Dillon CF, Hirsch R, Rasch EK, Gu Q (2007) Symptomatic hand osteoarthritis in the United States: prevalence and functional impairment estimates from the third U.S. National Health and Nutrition Examination Survey, 1991–1994. Am J Phys Med Rehabil 86:12–21

Dionne CE, Dunn KM, Croft PR (2006) Does back pain prevalence really decrease with increasing age? A systematic review. Age Ageing 35:229–234

Dionne CE, Dunn KM, Croft PR, Nachemson AL, Buchbinder R, Walker BF, Wyatt M, Cassidy JD, Rossignol M, Leboeuf-Yde C, Hartvigsen J, Leino-Arjas P, Latza U, Reis S, Gil del Real MT, Kovacs FM, Öberg B, Cedraschi C, Bouter LM, Koes BW, Picavet HSJ, van Tulder MW, Burton K, Foster NE, Macfarlane GJ, Thomas E, Underwood M, Waddell G, Shekelle P, Volinn E, von Korff M (2008) A consensus approach toward the standardization of back pain definitions for use in prevalence studies. Spine 33:95–103

Duncan EL, Brown MA (2008) Genetic studies in osteoporosis – the end of the beginning. Arthritis Res Ther 10:214. http://arthritis-research.com/content/10/5/214

Espallargues M, Sampietro-Colom L, Estrada MD, Solà M, del Río L, Setoain J (2001) Identifying bone-mass-related risk factors for fracture to guide bone densitometry measurements: a systematic review of the literature. Osteoporos Int 12:811–822

Essendrop M, Maul I, Läubli T, Riihimäki H, Schibye B (2002) Measures of low back function: a review of reproducibility studies. Clin Biomech 17:235–249

Fan ZJ, Silverstein BA, Bao S, Bonauto DK, Howard NL, Spielholz PO, Smith CK, Polissar NL, Viikari-Juntura E (2009) Quantitative exposure-response relations between physical workload and prevalence of lateral epicondylitis in a working population. Am J Ind Med 52:479–490

Felson DT (conference chair) (2000) Osteoarthritis: new insights. Part 1: the disease and its risk factors. Ann Intern Med 133:635–646

Felson DT (2009) Developments in the clinical understanding of osteoarthritis. Arthritis Res Ther 11:203. http://arthritits-research.com/content/11/1/203

Felson DT, Zhang Y, Hannan MT, Naimark A, Weissman B, Aliabadi P, Levy D (1997) Risk factors for incident radiographic knee osteoarthritis in the elderly. Arthritis Rheum 40:728–733

Felson DT, Niu J, Clancy M, Sack B, Alliabadi P, Zhang Y (2007) Effect of recreational physical activities on the development of knee osteoarthritis in older adults of different weights: the framingham study. Arthritis Rheum 57:6–12

Ferguson SA, Marras WS (1997) A literature review of low back disorder surveillance measures and risk factors. Clin Biomech 12:211–226

Gelber AC, Hochberg MC, Mead LA, Wang NY, Wigley FM, Klag M (2000) Joint injury in young adults and risk for subsequent knee and hip osteoarthritis. Ann Intern Med 133:321–328

Griffith LE, Hogg-Johnson S, Cole DC, Krause N, Hayden J, Burdorf A, Leclerc A, Coggon D, Bongers P, Walter SD, Shannon HS on behalf of the Meta-Analysis of Pain in the Lower Back and Work Exposures (MAPLE) Collaborative Group (2007) Low-back pain definitions in occupational studies were categorized for a meta-analysis using Delphi consensus methods. J Clin Epidemiol 60:625–633

Griffith LE, Wells RP, Shannon HS, Walter SD, Cole DC, Hogg-Johnson S on behalf of the Meta-Analysis of Pain in the Lower Back and Work Exposures (MAPLE) Collaborative Group (2008) Developing common metrics of mechanical exposures across aetiological studies of low back pain in working populations for use in meta-analysis. Occup Environ Med 65:467–481

Guh DP, Zhang W, Bansback N, Amarsi Z, Birmingham L, Anis AH (2009) The incidence of comorbidities related to obesity and overweight: a systematic review and meta-analysis. BMC Public Health 9:88. http://www.biomedcentral.com/1471-2458/9/88

Ha C, Roquelaure Y, Leclerc A, Touranchet A, Goldberg M, Imbernon E (2009) The French musculoskeletal disorders surveillance program: pays de la Loire network. Occup Environ Health 66:471–479

Haara M, Manninen P, Kröger H, Arokoski JPA, Kärkkäinen A, Knekt P, Aromaa A, Heliövaara M (2003) Osteoarthritis of finger joints in finns aged 30 or over: prevalence, determinants, and association with mortality. Ann Rheum Dis 62:151–158

Hadler NM, Gillings DB, Imbus HR, Levitin PM, Makuc D, Utsinger PD, Yount WJ, Slusser D, Moskowitz N (1978) Hand structure and function in an industrial setting. Arthritis Rheum 21:210–220

Hamberg-van Reenen HH, Ariens GA, Blatter BM, van Mechelen W, Bongers PM (2007) A systematic review of the relation between physical capacity and future low back and neck/shoulder pain. Pain 130:93–107

Harkness EF, Macfarlane GJ, Nahit ES, Silman AJ, McBeth J (2003) Risk factors for new-onset low back pain amongst cohorts of newly employed workers. Rheumatology 42:959–968

Harreby M, Kjer J, Hesselsoe G, Neergaard K (1996) Epidemiological aspects and risk factors for low back pain in 38-year-old men and women: a 25-year prospective cohort study of 640 school children. Eur Spine J 5:312–318

Hart DJ, Spector TD (1955) The classification and assessment of osteoarthritis. Bailliéres Clin Rheumatol 9:407–432

Hart DJ, Spector TD (2000) Definition and epidemiology of osteoarthritis of the hand: a review. Osteoarthritis Cartilage 8(Suppl A):S2–S7

Hartvigsen J, Lings S, Leboeuf-Yde C, Bakketeig L (2004) Psychosocial factors at work in relation to low back pain and consequences of low back pain; a systematic, critical review of prospective cohort studies. Occup Environ Med 61:e2

Hartvigsen J, Pedersen HC, Frederiksen H, Christensen K (2005) Small effect of genetic factors on neck pain in old age. A study of 2,108 Danish twins 70 years of age. Spine 30:206–208

Heistaro S, Vartiainen E, Heliövaara M, Puska P (1998) Trends of back pain in eastern Finland, 1972–1992, in relation to socioeconomic status and behavioral risk factors. Am J Epidemiol 48:671–682

Heliövaara M, Knekt P, Aromaa A (1987) Incidence and risk factors of herniated lumbar intervertebral disc or sciatica leading to hospitalization. J Chronic Dis 40:251–258

Heliövaara M, Mäkelä M, Knekt P, Impivaara O, Aromaa A (1991) Determinants of sciatica and low-back pain. Spine 16:608–614

Heliövaara M, Mäkelä M, Sievers K (1993) Tuki- ja liikuntaelinten sairaudet Suomessa. (Musculoskeletal diseases in Finland). Publication of the Social Insurance Institution AL, 35, Helsinki

Hestbaek L, Leboeuf-Yde C, Kyvik KO, Manniche C (2006) The course of low back pain from adolescence to adulthood: eight-year follow-up of 9600 twins. Spine 31:468–472

Hillman M, Wright A, Rajaratnam G, Tennant A, Chamberlain MA (1996) Prevalence of low back pain in the community: implications for service provision in Bradford, UK. J Epidemiol Community Health 50:347–352

Hogg-Johnson S, van der Velde G, Carroll LJ, Holm L, Cassidy JD, Guzman J, Côté P, Haldeman S, Ammendolia C, Carragee EJ, Hurwitz EL, Nordin M, Peloso PM (2008) The burden and determinants of neck pain in the general population. Results of the bone and joint decade 2000–2010 task force on neck pain and its associated disorders. Spine 33(4S):S39–S51

Hoogendoorn WE, van Poppel MNM, Bongers PM, Koes BW, Bouter LM (1999) Physical load during work and leisure time as risk factors for back pain. Scand J Work Environ Health 25:387–403

Hoogendoorn WE, Bongers PM, de Vet HCW, Douwes M, Koes BW, Miedema MC, Ariëns GAM, Bouter LM (2000a) Flexion and rotation of the trunk and lifting at work are risk factors for low back pain: results of a prospective cohort study. Spine 25:3087–3092

Hoogendoorn WE, van Poppel MNM, Bongers PM, Koes BW, Bouter LM (2000b) Systematic review of psychosocial factors at work and private life as risk factors for back pain. Spine 25:2114–2125

Jeffries LJ, Milanese SF, Grimmer-Somers KA (2007) Epidemiology of adolescent spinal pain: a systematic overview of the research literature. Spine 32:2630–2637

Jensen LK (2008a) Knee osteoarthritis: influence of work involving heavy lifting, kneeling, climbing stairs or ladders, or kneeling/squatting combined with heavy lifting. Occup Environ Med 65:72–89

Jensen LK (2008b) Hip osteoarthritis: influence of work with heavy lifting, climbing stairs or ladders, or combining kneeling/squatting with heavy lifting. Occup Environ Med 65:6–19

Jordan KM, Cooper C (2002) Epidemiology of osteoporosis. Best Pract Res Clin Rheumatol 16:795–806

Järvholm B, Lewold S, Malchau H, Vingård E (2005) Age, body weight, smoking habits and the risk of severe osteoarthritis in the hip and knee in men. Eur J Epidemiol 20:537–542

Järvholm B, From C, Lewold S, Malchau H, Vingård E (2008) Incidence of surgically treated osteoarthritis in the hip and knee in male construction workers. Occup Environ Health 65:275–278

Kaila-Kangas L (ed) (2007) Musculoskeletal disorders and diseases in Finland. Results of the Health 2000 Survey. Publications of the National Public Health Institute (B 25/2007), Helsinki http://www.ktl.fi/attachments/suomi/julkaisut/julkaisusarja_b/2007/2007b25.pdf

Kalichmann L, Hunter DJ (2008a) The genetics of intervertebral disc degeneration. Familial predisposition and heritability estimation. Joint Bone Spine 75:383–387

Kalichmann L, Hunter DJ (2008b) The genetics of intervertebral disc degeneration. Associated genes. Joint Bone Spine 75:388–396

Karasek R, Brisson C, Kawakami N, Houtman I, Bongers P, Amick B (1998) The job content questionnaire (JCQ): an instrument for internationally comparative assessment of psychosocial job characteristics. J Occup Health Psychol 3:322–355

Kauppila LI (2009) Atherosclerosis and disc degeneration/low back pain – a systematic review. Eur J Vasc Endovasc Surg 37:661–670

Kellgren JH (1963) The epidemiology of chronic rheumatism. Atlas of standard radiographs, vol 2. Blackwell Scientific, Oxford

Kellgren JH, Lawrence JS (1957) Radiological assessment of osteoarthritis. Ann Rheum Dis 16:494–501

Konstantinou K, Dunn KM (2008) Sciatica: review of epidemiological studies and prevalence estimates. Spine 33:2464–2472

Kuorinka I, Jonsson B, Kilbom Å, Vinterberg H, Biering-Sorensen F, Andersson G, Jorgensen K (1987) Standardised Nordic questionnaire for the analysis of musculoskeletal symptoms. Appl Ergon 18:233–237

Kurppa K, ViikariJuntura E, Kuosma E, Huuskonen M, Kivi P (1991) Incidence of tenosynovitis or peritendinitis and epicondylitis in a meatprocessing factory. Scand J Work Environ Health 17:32–37

Lagerstrom M, Hansson T, Hagberg M (1998) Work-related low-back problems in nursing. Scand J Work Environ Health 24:449–464

Lau EM, Egger P, Coggon D, Cooper C, Valenti L, O'Connell D (1995) Low back pain in Hong Kong: prevalence and characteristics compared with Britain. J Epidemiol Community Health 49:492-494

Lawrence JS (1969) Disc degeneration. Its frequency and relationship to symptoms. Ann Rheum Dis 28:121–138

Lawrence JS, Bremner JM, Bier F (1966) Osteo-arthrosis. Prevalence in the population and relationship between symptoms and X-ray changes. Ann Rheum Dis 25:1–23

Lawrence RC, Felson DT, Helmick CG, Arnold LM, Choi H, Deyo RA, Gabriel S, Hirsch R, Hochberg MC, Hunder GG, Jordan JM, Katz JN, Kremers HM, Wolfe F; National Arthritis Data Workgroup (2008) Estimates of the prevalence of arthritis and other rheumatic conditions in the United States. Arthr Rheum 58:26–35

Leboeuf-Yde C, Klougart N, Lauritzen T (1996) How common is low back pain in the Nordic population? Data from a recent study on a middle-aged general Danish population and four surveys previously conducted in the Nordic Countries. Spine 21:1518–1525

Lehto TU, Rönnemaa TE, Aalto TV, Helenius HYM (1990) Roentgenological arthrosis of the hand in dentists with reference to manual function. Community Dent Oral Epidemiol 18:37–41

Leino P, Berg M-A, Puska P (1994) Is back pain increasing? Results from national surveys in Finland during 1978/1979–1992. Scand J Rheumatol 23:269–276

Leino-Arjas P, Kaila-Kangas L, Notkola V, Keskimäki I, Mutanen P (2002) Inpatient hospital care for back disorders in relation to industry and occupation in Finland. Scand J Work Environ Health 28:304–313

Li G, Buckle P (1999) Current techniques for assessing physical exposure to work-related musculoskeletal risks, with emphasis on posture- based methods. Ergonomics 42:674–695

Linton SJ (2001) Occupational and psychological factors increase the risk for back pain: a systematic review. J Occup Rehab 11:53–66

Louw QA, Morris LD, Grimmer-Somers K (2007) The prevalence of low back pain in Africa: a systematic review. BMC Musculoskelet Disord 8:105

Macfarlane GJ, McBeth J, Garrow A, Silman AJ (2000) Life is as much a pain as it ever was. BMJ 321:897

Macfarlane GJ, Pallewatte N, Paudyal P, Blyth FM, Coggon D, Crombez G, Linton S, Leino-Arjas P, Silman AJ, Smeets RJ, van der Windt D (2009) Evaluation of work-related psychosocial factors and regional musculoskeletal pain: results from a EULAR Task Force. Ann Rheum Dis 68: 885–891

MacGregor AJ, Andrew T, Sambrook PN, Spector TD (2004) Structural, psychological, and genetic influences on low back and neck pain: a study of adult female twins. Arthritis Rheum 51:160–167

MacGregor AJ, Li Q, Spector TD, Williams FMK (2009) The genetic influence on radiographic osteoarthritis is site specific at the hand, hip and knee. Rheumatol 48:277–280

Manninen P, Heliövaara M, Riihimäki H, Suomalainen O (2002) Physical workload and risk of severe knee osteoarthritis. Scand J Work Environ Health 28:25–31

Martyn-St James M, Carroll S (2008) Meta-analysis of walking for preservation of bone mineral density in postmenopausal women. Bone 43:521–531

Martyn-St James M, Carroll S (2009) A meta-analysis of impact exercise on postmenopausal bone loss: the case for mixed loading exercise programmes. Br J Sports Med 43:898–908

Mason V (ed) (1994) The prevalence of back pain in Great Britain. A report on OPCS Omnibus Survey Data produced on behalf of the Department of Health. Her Majesty's Stationary Office, London, p.3

Mathiassen SE, Winkel J (1991) Quantifying variation in physical load using exposure-vs-time data. Ergonomics 34:1455–1468

May S, Littlewood C, Bishop A (2006) Reliability of procedures used in the physical examination of non-specific low back pain: a systematic review. Aust J Physiother 52:91–102

McBeth J, Jones K (2007) Epidemiology of chronic musculoskeletal pain. Best Pract Res Clin Rheumatol 21:403–425

Melchior M, Roquelaure Y, Evanoff B, Chastang J-F, Ha C, Imbernon E, Goldberg M, Leclerc A, and the Pays de la Loire Study Group (2006) Why are manual workers at high risk of upper limb disorders? The role of physical work factors in a random sample of workers in France. Occup Environ Med 63:754–761

Melton LJ III (2003) Epidemiology worldwide. Endocrinol Metab Clin North Am 32:1–13

Miranda H, Viikari-Juntura E, Heistaro S, Heliövaara M, Riihimäki H (2005) A population study on differences in the determinants of a specific shoulder disorder versus nonspecific shoulder pain without clinical findings. Am J Epidemiol 161:847–855

Miranda H, Punnett L, Viikari-Juntura E, Heliövaara M, Knekt P (2008) Physical work and chronic shoulder disorder. Results of a prospective population-based study. Ann Rheum Dis 67: 218–223

Mortimer M, Hjelm EW, Wiktorin C, Pernold G, Kilbom Å, Vingård E (1999) Validity of self-reported duration of work postures obtained by interview. MUSIC-Norrtälje Study Group. Appl Ergon 30:477–486

Moayyeri A (2008) The association between physical activity and osteoporotic fractures: a review of the evidence and implications for future research. Ann Epidemiol 18:827–35

Nachemson AL (2000) Introduction. In: Nachemson AL, Jonsson E (eds) Neck and back pain. The scientific evidence of causes, diagnosis, and treatment. Lippincott Williams & Wilkins, Philadelphia, pp 1–12

National Research Council and the Institute of Medicine (2001) Musculoskeletal disorders and the workplace: low back and upper extremities. Panel on musculoskeletal disorders and the workplace. Commission on behavioral and social sciences and education. National Academy Press, Washington, DC

Nordstrom DL, DeStefano F, Vierkant RA, Layde PM (1998) Incidence of diagnosed carpal tunnel syndrome in a general population. Epidemiology 9:342–345

Norlund AI, Waddell G (2000) Cost of back pain in some OECD countries. In: Nachemson AL, Jonsson E (eds) Neck and back pain. The scientific evidence of causes, diagnosis, and treatment. Lippincott Williams & Wilkins. Philadelphia, pp 421–425

O'Neill TW, Felsenberg D, Varlow J, Cooper C, Kanis JA, Silman AJ (1996) The prevalence of vertebral deformity in European men and women: the European vertebral osteoporosis study. J Bone Miner Res 11:1010–1018

Pafumi C, Chiarenza M, Zizza G, Roccasalva L, Ciotta L, Farina M, Pernicone G, Russo A, Maggi I, Bandiera S, Giardina P, Cavallaro A, Cianci A (2002) Role of DEXA and ultrasonometry in the evaluation of osteoporotic risk in postmenopausal women. Maturitas 42:113–117

Palmer KT, Walsh K, Bendall H, Cooper C, Coggon D (2000) Back pain in Britain: comparison of two prevalence surveys at an interval of 10 years. BMJ 320:1577–1578

Powell MC, Wilson M, Szypryt P, Symonds EM, Worthington BS (1986) Prevalence of lumbar disc degeneration observed by magnetic resonance in symptomless women. Lancet i:1366–1367

Pfirrmann CWA, Metzdorf A, Zanetti M, Hodler J, Boos N (2001) Magnetic resonance classification of lumbar intervertebral disc degeneration. Spine 26:1873–1878

Punnett L, Prüss-Üstün A, Nelson DI, Fingerhut MA, Leigh J, Tak SW, Phillips S (2005) Estimating the global burden of low back pain attributable to combined occupational exposures. Am J Ind Med 48:459–469

Rafnsson V, Steingrimsdottir OA, Olafsson MH, Sveinsdottir T (1989) Musculoskeletal disorders in the Icelandic population (in Swedish). Nord Med 104:104–107

Raspe H, Matthis C, Croft P, O'Neill T, European Vertebral Osteoporosis Group (2004) Variation in back pain between countries: the example of Britain and Germany. Spine 29:1017–1021

Reijman M, Hazes JMW, Koes BW, Verhagen AP, Bierma-Zeinstra SMA (2004) Validity, reliability, and applicability of seven definitions of hip osteoarthritis used in epidemiological studies: a systematic appraisal. Ann Rheu Dis 63:226–232

Riihimäki H, Viikari-Juntura E (2000) Back and limb disorders. In: McDonald C (ed) Epidemiology of work-related diseases, 2nd edn. BMJ Publishing Group, Bristol, pp 233–265

Riihimäki H, Heliövaara M and the working group for musculoskeletal diseases (2004) Musculoskeletal diseases. In Aromaa A, Koskinen S (eds) Health and functional capacity in Finland. Baseline results of the Health 2000 Health Examination Survey. Publications of the National Public Health Institute (B12/2004), Helsinki, pp 55–58. http://www.ktl.fi/attachments/suomi/julkaisut/julkaisusarja_b/2004b12.pdf

Roquelaure Y, Ha C, Leclerc A, Touranchet A, Sauteron M, Melchior M, Imbernon E, Goldberg M (2006) Epidemiologic surveillance of upper-extremity musculoskeletal disorders in the working population. Arthritis Rheum 55:765–778

Salminen JJ, Erkintalo MO, Pentti J, Oksanen A, Kormano MJ (1999) Recurrent low back pain and early disc degeneration in the young. Spine 24:1316–1321

Schiphof D, de Klerk BM, Koes BW, Bierma-Zeinstra S (2008) Good reliability, questionable validity of 25 different classification criteria of knee osteoarthritis: a systematic appraisal. J Clin Epidemiol 61:1205–1215

Sharma L (2001) Local factors in osteoarthritis. Curr Opin Rheumatol 13:441–446

Shiri R, Viikari-Juntura E, Varonen H, Heliövaara M (2006) Prevalence and determinants of lateral and medial epicondylitis: a population study. Am J Epidemiol 164:1065–1074

Shiri R, Karppinen J, Leino-Arjas P, Solovieva S, Varonen H, Kalso E, Ukkola O, Viikari-Juntura E (2007) Cardiovascular and lifestyle factors in lumbar radicular pain or clinically defined sciatica: a systematic review. Eur J Spine 16:2043–2054

Shiri R, Miranda H, Heliövaara M, Viikari-Juntura E (2009) Physical work load factors and carpal tunnel syndrome: a population-based study. Occup Environ Med 66:368–373

Shiri R, Karppinen J, Leino-Arjas P, Solovieva S, Viikari-Juntura E (2010a) The association between obesity and low back pain: a meta analysis. Am J Epidemiol 171:135–154

Shiri R, Karppinen J, Leino-Arjas P, Solovieva S, Viikari-Juntura E (2010b) The association between smoking and low back pain: a meta analysis. Am J Med 123:87.e7–87.e35

Silverstein B, Viikari-Juntura E, Kalat J (2002) Use of a prevention index to identify industries at high risk for work-related musculoskeletal disorders of the neck, back, and upper extremity in Washington state, 1990–1998. Am J Ind Med 41:149–169

Silverstein B, Viikari-Juntura E, Fan ZJ, Bonauto DK, Bao S, Smith C (2006) Natural course of nontraumatic rotator cuff tendinitis and shoulder symptoms in a working population. Scand J work Environ Health 32:99–108

Silverstein BA, Bao SS, Fan ZJ, Howard N, Smith C, Spielholz P, Bonauto D, Viikari-Juntura E (2008) Rotator cuff syndrome: personal, work-related psychosocial and physical load factors. J Occup Environ Med 50:1062–1076

Sluiter JK, Rest KM, Frings-Dresen MHW (2001) Criteria document for evaluating the work-relatedness of upper extremity musculoskeletal disorders. Scand J Work Environ Health 27(Suppl 1):1–102

Solovieva S, Vehmas T, Riihimäki H, Luoma K, Leino-Arjas P (2005) Hand use and patterns of joint involvement in osteoarthritis. A comparison of female dentists and teachers. Rheumatology 44:521–528

Solovieva S, Vehmas T, Riihimäki H, Takala E-P, Murtomaa H, Luoma K, Leino-Arjas P (2006) Finger osteoarthritis and differences in dental work tasks. J Dent Res 85:344–348

Sowers M (2001) Epidemiology of risk factors for osteoarthritis: systemic factors. Curr Opin Rheumatol 13:447–451

Spector T, MacGregor AJ (2004) Risk factors for osteoarthritis: genetics. Osteoarthritis Cartilage 12:S39–S44

Stanton TR, Latimer J, Maher CG, Hancock M (2009) Definitions of recurrence of an episode of low back pain. A systematic review. Spine 43:E316–E322

Tanamas S, Hanna FS, Cicuttini FM, Wluka AE, Berry P, Urquhart DM (2009) Does knee malalignment increase the risk of development and progression of knee osteoarthritis? A systematic review. Arthritis Rheum 61:459–467

Tang BMP, Eslick GD, Nowson C, Smith C, Bensoussan A (2007) Use of calcium or calcium in combination with vitamin D supplementation to prevent fractures and bone loss in people aged 50 years and older: a meta-analysis. Lancet 379:657–666

Tegeder I, Lötsch J (2009) Current evidence for a modulation of low back pain by human genetic variants. J Cell Mol Med 13:1605–1619

Valdes AM, Doherty M, Spector TD (2008) The additive effect of individual genes in predicting risk of knee osteoarthritis. Ann Rheum Dis 67:124–127

van der Beek AJ, Frings-Dresen MHW (1998) Assessment of mechanical exposure in ergonomic epidemiology. Occup Environ Med 55:291–299

van der Beek AJ, Driessen MT, Proper KI, van Tulder MW, Anema JR, Bongers PM (2009) The effectiveness of ergonomic workplace interventions on low back pain and neck pain; a systematic review. In: Proceedings of the 17th World Congress on Ergonomics, IEA 2009 International Ergonomics Association, Peking, 9–14 Aug 2009

van Saase JLCM, van Romunde LKJ, Vandenbroucke JP, Valkenburg HA (1989) Epidemiology of osteoarthritis: zoetermeer survey. Comparison of radiological osteoarthritis in a Dutch population with that in 10 other populations. Ann Rheum Dis 48:271–280

van Staa TP, Dennison EM, Leufkens HGM, Cooper C (2001) Epidemiology of fractures in England and Wales. Bone 29:517–522 van Tulder MW, Assendelft WJJ, Koes BW, Bouter LM (1997) Spinal radiographic findings and nonspecific low back pain. A systematic review of observational studies. Spine 22:427–434

Viikari-Juntura E, Rauas S, Martikainen R, Kuosma E, Riihimäki H, Takala E-P, Saarenmaa K (1996) Validity of self-reported physical work load in epidemiologic studies on musculoskeletal disorders. Scand J Work Environ Health 22:251–259

Vingård E, Alfredsson L, Fellenius E, Goldie I, Hogstedt C, Köster M (1991a) Coxarthrosis and physical load from occupation. Scand J Work Environ Health 17:104–109

Vingård E, Alfredsson L, Goldie I, Hogstedt C (1991b) Occupation and osteoarthrosis of the hip and knee: a register-based cohort study. Int J Epidemiol 20:1025–1031

Vingård E, Alfredsson L, Hagberg M, Kilbom A, Theorell T, Waldenstrom M, Hjelm EW, Wiktorin C, Hogstedt C (2000) To what extent do current and past physical and psychosocial occupational factors explain care-seeking for low back pain in a working population? Results from the musculoskeletal intervention center-norrtalje study. Spine 15:493–500

Vingård E (2001a) Carpal tunnel syndrome (in Swedish with English summary) In: Hansson T, Westerholm P (eds) Arbete och besvär i rörelseorganen. En vetenskaplig värdering av frågor om samband. National Institute of Working Life, Stockholm. Arbete och Hälsa 12:173–183

Vingård E (2001b) Epicondylitis and work (in Swedish with English summary) In: Hansson T, Westerholm P (eds) Arbete och besvär i rörelseorganen. En vetenskaplig värdering av frågor om samband. National Institute of Working Life, Stockholm. Arbete och Hälsa 12:161–171

von Korff M, Dworkin SF, Le Resche L (1990) Graded chronic pain status: an epidemiologic evaluation. Pain 40:270–291

Walker BF (2000) The prevalence of low back pain: a systematic review of the literature from 1966 to 1998. J Spin Disorders 13:205–217

Walker-Bone K, Walter G, Cooper C (2002) Recent developments in the epidemiology of osteoporosis. Curr Opin Rheumatol 14:411–415

Walker-Bone K, Palmer KT, Reading I, Coggon D, Cooper C (2004) Prevalence and impact of musculoskeletal disorders of the upper limb in the general population. Arthritis Rheum 51: 642–651

Walker-Bone K, Reading I, Coggon D, Cooper C, Palmer KT (2006) Risk factors for specific upper limb disorders as compared with non-specific upper limb pain: assessing the utility of a structured examination schedule. Occup Med 56:243–250

Walsh K, Cruddas M, Coggon D (1992) Low back pain in eight areas of Britain. J Epidemiol Community Health 46:227–230

Wells R, Norman R, Neumann P, Andrews D, Frank J, Shannon H (1997) Assessment of physical workload in epidemiologic studies: common measurement metrics for exposure assessment. Ergonomics 40:51–61

Welten DC, Kemper HCG, Post B, Staveren WA (1995) A meta-analysis of the effect of calcium intake on bone mass in young and middle aged females and males. J Nutr 125:2802–2813

Williams FMK, Spector TD (2006) Recent advances in the genetics of osteoporosis. J Musculoskelet Neurol Interact 6:27–35

Winkel J, Mathiassen SE (1994) Assessment of physical load in epidemiologic studies: concepts, issues, and operational considerations. Ergonomics 37:979–988

World Health Organisation (WHO) (1992) International statistical classification of diseases and related health problems, 1989 revision. WHO, Geneva

World Health Organisation (WHO) (1994) Assessment of fracture risk and its application to screening for postmenopausal women. WHO Technical Report Series 843. WHO, Geneva

Yusuf E, Nelissen R, Ioan-Facsinay A, Stojanovic-Susulic V, Degroot J, van Osch G, Middeldorp S, Huizinga T, Kloppenburg M (2010) Association between weight or body mass index and hand osteoarthritis: a systematic review. Ann Rheum Dis 69:761–765

Zhang Y, Niu J, Kelly-Hayes M, Chaisson CE, Alibadi P, Felson DT (2002) Prevalence of symptomatic hand osteoarthritis and its impact on functional status among the elderly: the Framingham Study. Am J Epidemiol 156:1021–1027

Epidemiology of Obesity

55

Brian K. Kit, Cynthia L. Ogden, and Katherine M. Flegal

Contents

B. K. Kit (✉)
Division of Health and Nutrition Examination Surveys, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, USA

Epidemic Intelligence Service, Scientific Education and Professional Development Program Office, Centers for Disease Control and Prevention, Hyattsville, MD, USA

C. L. Ogden
Division of Health and Nutrition Examination Surveys, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, USA

K. M. Flegal
Office of the Director, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_55,

55.1 Introduction

Obesity results from an imbalance between energy intake and energy expenditure. Factors that influence energy intake include the quantity and quality of foods and beverages consumed, while physical activity levels influence energy expenditure. Energy balance is further influenced by genetic, metabolic, and environmental factors.

Today, obesity affects individuals around the globe. However, the prevalence of obesity is not equally distributed. For example, there are variations in obesity by country and within countries by geographical region, socioeconomics, and demographic variables including race/ethnicity, gender, and age.

Understanding the epidemiological features of obesity is an important step in determining and prioritizing interventions. To this end, this chapter will provide an overview of measures of body composition, definitions of obesity, and prevalence of obesity. Further, there will be a brief discussion of factors that contribute to obesity.

55.2 Body Composition

Adiposity refers to body fat and obesity to an excess of body fat. Fat is stored in the body as adipose tissue and to a lesser degree lipids. The majority of adipose tissue is located subcutaneously (under the skin) and surrounding the viscera (organs including stomach, liver, and heart). An individual's weight, on the other hand, is determined by a combination of adipose and lean body mass. Lean body mass includes muscle, organs, bone, and extracellular water.

One of the challenges in obesity research is discerning the relative contribution of an individual's total body fat to his/her total weight. There are several available methods to measure body composition, and each has its own strengths and limitations in its ability to assess adiposity. These methods have been extensively discussed in nutritional textbooks (Willett 1988; Hu 2008) and will be discussed briefly in this chapter.

55.3 Measures of Body Composition

55.3.1 Hydrodensitometry

Hydrodensitometry also known as underwater weighing, is considered the gold standard for measuring body fatness. Hydrodensitometry is based on the principle that fat is lighter than other body constituents (Willett 1988). The procedure involves weighing individuals in the air, and then again after they are submerged underwater. Following the procedure, calculation of body volume, body density, and percent body fat can be performed (Brozek et al. 1963). Although still the gold standard, hydrodensitometry is used infrequently except in research settings and to assess the validity of other measures (Hu 2008). The cost, time, and requirements for significant cooperation of individuals being weighed are limitations to its use.

55.3.2 Dual Energy X-Ray Absorptiometry (DXA)

DXA uses an x-ray beam combined with a whole body scanner to measure both bone mass and soft tissue composition (Willett 1988). DXA is able to distinguish fat and lean body mass by the differential absorption of x-rays by these tissues (Willett 1988) and can provide whole body or regional estimates. The reproducibility (Lohman 1981; Cordero-MacIntyre et al. 2002), accuracy, and precision (Kiebzak et al. 2000) of DXA in measuring total body fat has increased its use in obesity research settings (Hu 2008). Although there is radiation involved, the dose is sufficiently low to allow its use in all age groups, but its use is avoided during pregnancy. For example, the radiation exposure from a DXA scan is roughly equivalent to an airline flight from Los Angeles, California, to New York, New York. The variability in estimates between machines from different manufacturers should be considered if comparisons are being made (Covey et al. 2010). The expense associated with DXA limits its use in both research and clinical settings.

55.3.3 Bioelectric Impedance Analysis (BIA)

BIA is a relatively low-cost method for estimating percent body fat by measuring the impedance or resistance to an electrical current as it travels though the body (Hu 2008). Incorporating sex, age, and race into models to predict body fat improves the validity of BIA (Sun et al. 2003). However, level of body fatness itself may influence the BIA estimates, with an overestimation of body fat for healthy weight individuals but underestimation of body fat for obese individuals (Sun et al. 2005; Fakhrawi et al. 2009; Savastano et al. 2010). Further, some evidence suggests that body mass index (BMI) provides as much information for predicting adiposity as BIA (Willett et al. 2006). Other factors impacting BIA estimates include hydration status and body shape (Hu 2008). Different BIA devices employ different measurement techniques which impact body fat estimates (Pateyjohns et al. 2006).

55.3.4 Anthropometry

Weight and height measurements, often combined into BMI, are widely used in epidemiological obesity studies. However, other measurements such as waist circumference and skinfold thicknesses are also commonly used.

55.3.4.1 Body Mass Index (BMI)

BMI is calculated as the weight in kilograms divided by the square of height in meters. Lean body mass, in addition to fat, contributes to an individual's weight and therefore BMI. BMI is correlated with body fatness measured by densitometry. However, there are differences by gender, race, and age in the ability of BMI to predict body fat (Gallagher et al. 1996). For example, women have lower muscle (Gallagher et al. 1997) and bone mass (Tuck et al. 2005) than men. Therefore, at the same BMI, women tend to have a higher percent body fat. Similarly, at the same BMI, compared with whites, Asians have more body fat (Deurenberg et al. 2002) and blacks have less body fat (Wagner and Heyward 2000; Flegal et al. 2010b). There are also changes in the fat mass with aging. Among children, there is an increase in fat mass during the first year of life followed by a period of stabilization and subsequent increase between the ages of 3 and 7 years (Rolland-Cachera et al. 1984). Among adults, there are age-related declines in lean body mass and increases in fat mass (Evans and Campbell 1993).

Despite its limitations, the use of BMI has been widely adopted in clinical practice as well as research settings to assess obesity.

Measured Versus Reported BMI

One of the advantages of BMI is its low cost for assessment of obesity when compared with other assessment tools. However, collecting height and weight data is not without associated costs. Self-reported data, rather than data measured by a trained professional, may be more feasible for some surveillance projects. Self-reported height and weight are subject to their own set of limitations which should be considered when evaluating and drawing conclusions from the data. Among adolescents and adults, individuals tend to underreport their weight and overreport their height, which may lead to an underestimation of obesity (Bostrom and Diderichsen 1997; Abraham et al. 2004; Gillum and Sempos 2005). For children, parent-reported height and weight are sometimes collected. In this situation, parents may underreport height among younger children (Akinbami and Ogden 2009).

55.3.4.2 Other Anthropometric Data

Waist circumference and skinfold thicknesses are two other common anthropometric measurements used to assess body fatness. Measurement sites for waist circumference include the natural waist (halfway between the lowest rib and iliac crest) or the level of the umbilicus (Wang et al. 2003). Some research suggests that waist circumference may be more strongly correlated with health outcomes than BMI (Janssen et al. 2004), particularly among Asians, while other research suggests the two have a similar relation to percent body fat (Flegal et al. 2009). The use of waist circumference is limited by the lack of standardization of measurement site and, for children, the lack of reference values. Skinfold measurements are collected using special calipers at specific locations using a standard technique (Hu 2008). Used in combination with prediction equations (Jackson and Pollock 1978), skinfold measurements provide body fat estimates (Peterson et al. 2003). However, measurement error associated with skinfold thickness limits its use in clinical and research settings (Bray et al. 1978; Marks et al. 1989; Ulijaszek and Kerr 1999).

55.3.5 Other Measures of Body Composition

Other measures of body composition include air displacement plethysmology (ADP) and radiographic imaging modalities including computed tomography (CT) and magnetic resonance imaging (MRI). ADP employs a similar technique to hydrodensitometry but does not require individuals to be submerged underwater as it uses air rather than water displacement for its calculations (Hu 2008). Body fat estimates from ADP are considered reliable (Fields et al. 2000; Noreen and Lemon 2006; Anderson 2007). Imaging modalities, including MRI and CT, allow for estimation of adipose tissue and quantification of the distribution of adipose tissue into visceral and subcutaneous deposits (van der Kooy and Seidell 1993; Plourde 1997). Thus, these images allow for fat estimates within individual organs as well as total body fat. However, the cost and inability of most MRI machines and CT scanners to accommodate very large individuals (BMI > 40 kilogram/meter^2) limit their application. Additionally, radiation exposure associated with a CT scan is a consideration (Brenner and Hall 2007).

55.4 Definitions for Obesity

Body mass index, or BMI, is the basis of the definitions for overweight and obesity (National Institutes of Health 1998; World Health Organization 2000) in population studies. Separate definitions are applied for children and adults. These definitions will be reviewed below.

55.4.1 Defining Overweight and Obesity: Adults

In adults, overweight and obesity are defined by fixed BMI values beyond which there are additional health risks, including hypertension, diabetes, and hypercholesterolemia. Obesity is defined similarly by the National Institutes of Health (NIH)/National Heart, Lung, and Blood Institute (NHLBI) in the United States (National Institutes of Health 1998) and by the World Health Organization (WHO) (World Health Organization 2000). Both organizations define obesity as a BMI of 30 kg/m^2 or greater. Cut-offs for obesity have been further divided into classes of obesity with class I corresponding to a BMI of 30–34.9 kg/m^2, class II to a BMI of 35–39.9 kg/m^2, and class III to a BMI greater than or equal to 40 kg/m^2. Class III obesity is sometimes referred to as severe or extreme obesity. NHLBI defines overweight as BMI greater than or equal to 25 kg/m^2 but less than 30 kg/m^2.

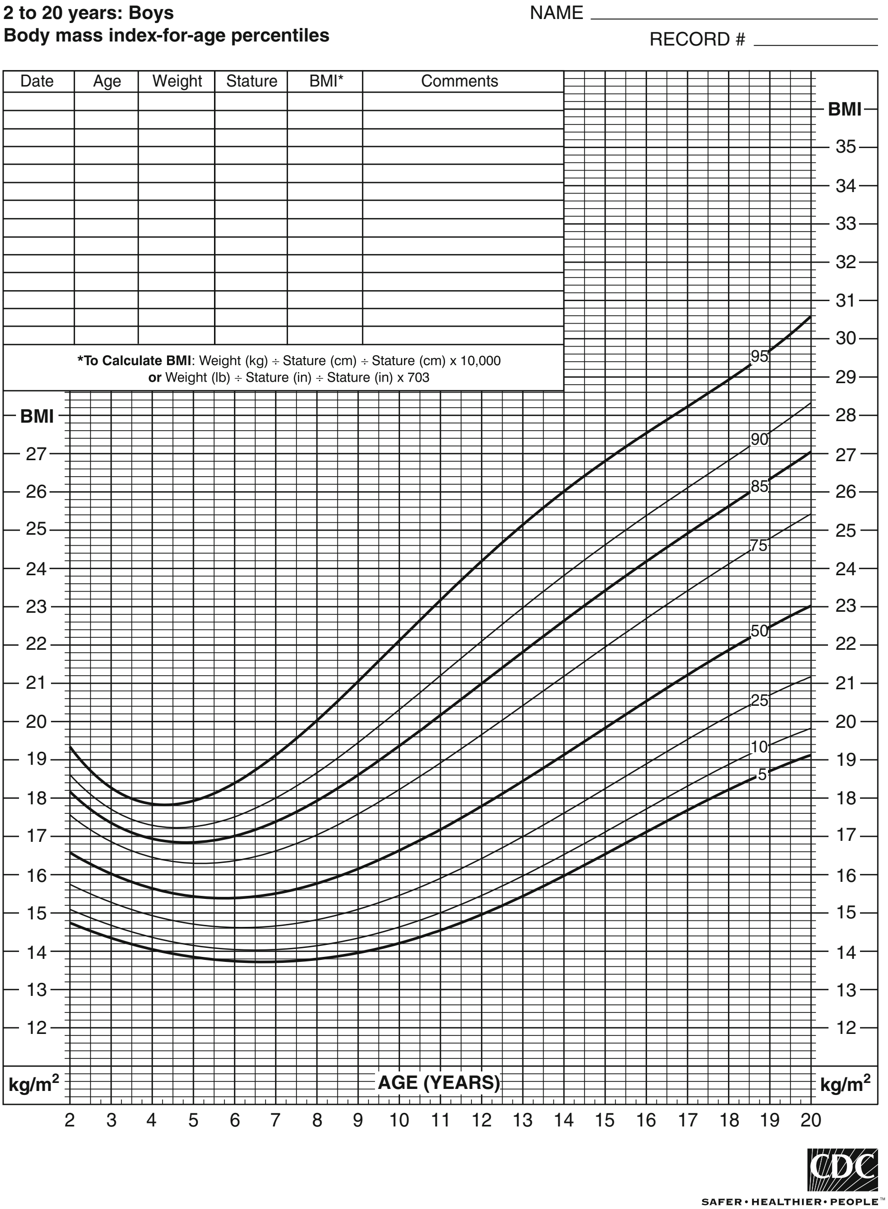
55.4.2 Defining Overweight and Obesity: Children

Unlike in adults, there are no single definitions of overweight and obesity for children. Health risks associated with obesity are not commonly detected during childhood. In the absence of such outcomes and because BMI varies with age, a statistical definition of overweight and obesity based on the percentile of BMI-for-age in a specified reference population is often used for children (Himes and Dietz 1994; Barlow and Dietz 1998; Krebs et al. 2007). The use of BMI *z*-score, based on standard deviations observed above the mean, after a normalizing transformation, is also used. Overweight and obesity in children are often defined as a BMI between the 85th and 95th percentiles and a BMI at or above the 95th percentile (Himes and Dietz 1994; Barlow and Dietz 1998; Krebs et al. 2007), respectively, or a BMI z-score between $+2$ and $+3$ and a BMI *z*-score greater than $+3$, respectively (de Onis and Blossner 2003; Flegal and Ogden 2011).

For children, where overweight and obesity are defined in reference to a specified population, the composition of the population to which the children are being compared is important to consider. In the United States (Kuczmarski et al. 2002) and other countries (Cole et al. 1998; Cacciari et al. 2006), the use of historical reference data drawn from a nationally representative sample of children from within the country is a common method by which children are compared in defining overweight and obesity. In the United States, the Centers for Disease Control and Prevention (CDC) 2000 growth charts (Kuczmarski et al. 2002) are commonly used as the reference for defining overweight and obesity. To develop the CDC growth charts, measured heights and weights from a representative sample of children were collected from national survey data in the 1960s, the 1970s, and, for children under 6 years, 1988–1994 (Kuczmarski et al. 2002). An example of the 2000 CDC growth charts is shown in Fig. 55.1, which demonstrates BMI-for-age for boys, ages 2–20 years.

Another reference to define overweight and obesity in children is the 2000 International Obesity Task Force (IOTF) published BMI cut-off values based on six nationally representative data sets with measured heights and weights from Brazil, Great Britain, Hong Kong, the Netherlands, Singapore, and the United States (Cole et al. 2000). These values, often referred to as the International Obesity Task Force cut-off values, were developed to provide a common definition for childhood overweight and obesity and thus facilitate comparisons of prevalence estimates to be used by researchers and policymakers in the international community (Cole et al. 2000). These values were not intended to replace national reference data. The cut-off values themselves were chosen as the percentiles that matched the adult cut-offs for overweight and obesity at age 18 years (i.e., a BMI of 25 kg/m^2 and a BMI of 30 kg/m^2, respectively).

In 2006, the World Health Organization (WHO) released BMI-for-age growth standards for preschool-age children from birth to 5 years of age. The approach for the development of these growth charts was different from that used for the CDC 2000 growth charts and for the IOTF cut-off values. Rather than using a descriptive reference sample, the WHO growth standards were based on data from children living in optimal conditions for growth in selected sites in Brazil, Ghana, India, Norway, Oman, and the United States of America (WHO Multicentre Growth Reference Study Group 2006). Specific inclusion criteria included no known health or environmental constraints to growth, exclusive or predominant breastfeeding for at least 4 months, continued breastfeeding to at least 12 months of age, no maternal smoking before and after delivery, single term birth, and absence of



2 to 20 years: Boys
Body mass index-for-age percentiles

NAME ______________________

RECORD # __________

Date	Age	Weight	Stature	BMI*	Comments

***To Calculate BMI**: Weight (kg) ÷ Stature (cm) ÷ Stature (cm) x 10,000
or Weight (lb) ÷ Stature (in) ÷ Stature (in) x 703

Fig. 55.1 Body mass index-for-age percentiles for boys, ages 2–20 years (Source: Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2000). http://www.cdc.gov/growthcharts, Published May 30, 2000 (modified 10/16/2000))

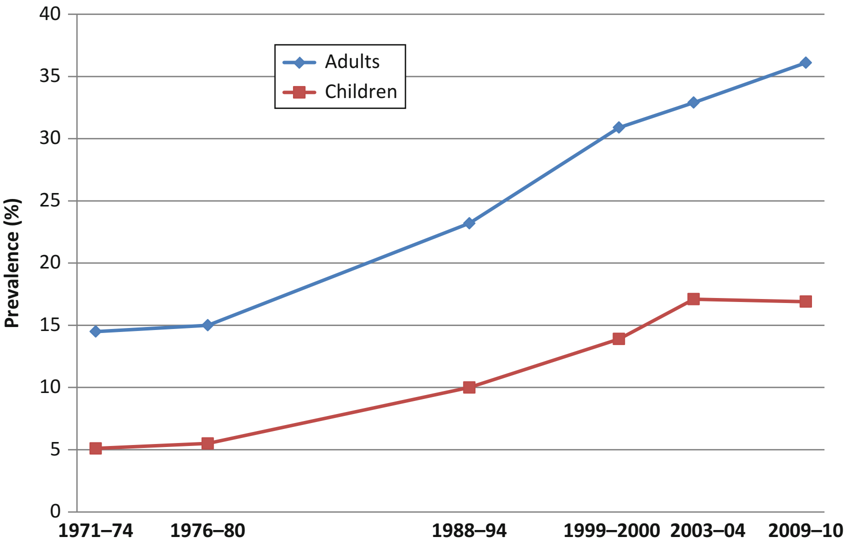


Fig. 55.2 Obesity trends among persons aged 2–19 and 20–74 years, USA, 1971–1974 through 2007–2008. Obesity defined as BMI >= 30 for adults and high BMI as BMI-for-age >= 95th percentile on the sex specific 2000 CDC Growth Charts for 2–19 years (Sources: Fryar et al. (2012a, b))

significant morbidity (de Onis et al. 2004). The WHO growth charts are intended to present a standard of physiologic growth. In 2007, the WHO released growth references for children and adolescents aged 5–19 years (de Onis et al. 2007). Unlike the growth standards for younger children, the WHO growth references for children and adolescents aged 5–19 years are based on national survey data from the United States.

55.5 Prevalence and Trends of Obesity

Temporal trends in obesity prevalence assessed by BMI for adults and children and adolescents in the United States between 1971 and 2010 are presented in Fig. 55.2 Fryar et al. (2012a, b). Generally, the prevalence of obesity in the United States increased between 1976–1980 and 1999–2000 for adults and children and adolescents. However, there are differences in obesity prevalence by demographic variables, including sex and race/ethnicity, which will be discussed below.

55.5.1 Prevalence and Trends of Obesity in the United States: Adults

Using the NHLBI obesity definition, between 1976–1980 and 1999–2000, the prevalence of obesity increased from 12.7% to 27.7% in men and 17.0–33.4% in

Table 55.1 Prevalence of obesity[a] among adults aged 20–74 years by sex, United States, 1971–2010[b]

	Sex	
Survey years[b]	Men(%)	Women(%)
1971–1974	12.1	16.6
1976–1980	12.7	17.0
1988–1994	20.5	25.9
1999–2000	27.7	34.0
2001–2002	28.3	34.1
2003–2004	31.7	34.0
2005–2006	33.8	36.3
2007–2008	32.5	36.2
2009–2010	35.9	36.1

[a] Obesity defined as BMI ≥ 30

[b] Based on data from the National Health and Nutrition Examination Surveys reported by Flegal et al. (2002, 2010a); Fryar et al. (2012b)

Table 55.2 Prevalence of obesity[a] among children aged 2–19 years by sex, United States, 1971–2010[b]

	Sex	
Survey years[b]	Boys(%)	Girls(%)
1971–1974	5.2	5.0
1976–1980	5.4	5.7
1988–1994	10.2	9.8
1999–2000	14.0	13.8
2001–2002	16.4	14.3
2003–2004	18.2	16.0
2005–2006	15.9	14.9
2007–2008	17.7	15.9
2009–2010	18.6	15.0

[a] Obesity defined as BMI ≥ 95th percentile on the 2000 sex- and age-specific CDC growth charts

[b] Based on data from the National Health and Nutrition Examination Surveys reported by Ogden et al. (2010), Moreno et al. (2011), and Fryar et al. (2012a)

women (Table 55.1). In 2005–2006, the prevalence of obesity was 33.3% for men and 35.3% for women. More recently, there has not been a significant increase for women; however, there may be increases in obesity prevalence for men (Flegal et al. 2010a).

55.5.2 Prevalence and Trends of Obesity in the United States: Children and Adolescents

Obesity, defined as a BMI ≥ 95th percentile on sex- and age-specific 2000 CDC growth charts, increased among boys from 5.4% to 14% between 1976–1980 and 1999–2000 (Table 55.2). Among girls, the increase was from 5.7% to 13.8%.

Of note, the estimates of childhood obesity in 1976–1980 are close to 5% by definition; data from this time period were included in the construction of the 2000 CDC growth charts and contributes to the statistical definition of obesity. The prevalence of obesity in 2005–2006 was 15.9% for boys and 15.0% for girls. More recent data suggests among girls that childhood obesity has remained stable in the United States since 2000.

55.5.3 Race/Ethnicity and Obesity in the United States

In the United States, there are differences in BMI-defined obesity by race/ethnicity. Hispanic boys, but not Hispanic girls, have a higher likelihood of obesity than non-Hispanic whites (Ogden et al. 2010). Among girls, non-Hispanic blacks have a higher likelihood of obesity, measured by BMI, than non-Hispanic whites (Ogden et al. 2010). Among men, obesity prevalence is similar across race/ethnic groups (Flegal et al. 2010a). Among women, non-Hispanic blacks and Mexican Americans have a higher likelihood of obesity than non-Hispanic whites (Flegal et al. 2010a). Describing racial and ethnic differences in the prevalence of obesity, defined by BMI, has some limitations as there are differences in body fat between race and ethnic groups at the same BMI level (Deurenberg et al. 2002; Flegal et al. 2010b).

55.5.4 Global Obesity Prevalence

The prevalence of obesity in other North American countries is generally lower than that seen in the United States. For example, in North American studies using measured heights and weights, the prevalence of obesity for adult men was 22.9% in Canada (in 2004) (Tjepkema 2006), 33.8% in the United States (in 2005–2006, Table 55.1), and 24.2% in Mexico (in 2006) (Olaiz-Fernàndez et al. 2006). Likewise among women, the prevalence of obesity is 23.2% in Canada (in 2004) (Tjepkema 2006), 36.3% in the United States (in 2005–2006, Table 55.1), and 34.5% in Mexico (in 2006) (Olaiz-Fernàndez et al. 2006). However, obesity extends beyond North America and is increasingly being referred to as a global epidemic (World Health Organization 2000). To track the global prevalence of adult obesity, the WHO maintains the Global Database on Body Mass Index (World Health Organization 2010). Although obesity prevalence estimates in the database may not be directly comparable because of differences in sampling procedures, sampling years, collection methods, and ages of inclusion, they provide a comprehensive overview of obesity prevalence from around the world. For example, Nauru is noted to have the highest obesity prevalence (78.5%), while Vietnam is noted to have the lowest prevalence of obesity (0.46%) (World Health Organization 2010).

In addition to the challenges in comparing adult obesity prevalence between countries, the challenge is further magnified for children because of differences in definition of obesity used in the studies. One study of children ages 11–15 years from 41 countries conducted in 2005–2006 defined obesity according to the IOTF

Table 55.3 Prevalence of obesity among children by country/region using the IOTF definition[a]

Country	Boys	Girls	Years	Ages
Cyprus	6.9	5.7	1999–2000	6–17
Great Britain	6.9	7.4	2007	2–10
	4.8	6.1	2007	11–18
France	3.8	3.6	2000	7–9
Beijing, China	8.0	3.2	2004	2–18
Canada	8.1	7.2	2004	2–17

[a] Based on measured height and weight as reported by Rolland-Cachera et al. (2002), Savva et al. (2005), Shields and Tremblay (2010), and Stamatakis et al. (2010)

cut-offs using self-reported height and weight (Haug et al. 2009). Results indicated wide ranges of obesity prevalence between countries, with a notably high prevalence in Malta (10.7%) and the United States (8.9%) and notably low prevalence in Latvia (0.9%), Lithuania (0.9%), and Russia (0.9%). Childhood obesity prevalence, using the IOTF definition, for select cities/countries based on measured data collected during a similar time frame is presented in Table 55.3. Like in the United States, the rate of increase of childhood obesity may have slowed in other countries, including Australia (Olds et al. 2010) and Sweden (Sjoberg et al. 2008).

55.6 Contributing Factors to Obesity: Diet

55.6.1 Temporal and Age-Related Trends in Food Availability and Mean Dietary Intakes

The prevalence of obesity in the United States and elsewhere increased during the 1980s and 1990s. During this time period, there were concomitant increases in the per capita food energy availability in the food supply (Balanza et al. 2007). For example, between 1970 and 1994, there was a 15.2% increase in the daily per capita availability of food energy in the United States (Harnack et al. 2000). Fats and oils, caloric sweeteners, carbonated soft drinks, and other food sources contributed to this increase (Putnam et al. 2002).

Among adults in the United States, the increased availability of food energy in the food supply may have translated into an increase in mean energy intake. For men, the mean daily energy intake (in kilocalories) increased from 2,450 to 2,618 kilocalories (kcal) between 1971–1974 and 1999–2000 (Wright et al. 2004). During the same time period, the increase for women was from 1,542 to 1,877 kcal (Wright et al. 2004). In addition to changes in caloric content of American's diet, there have been changes in its composition. For example, among the macronutrients, there was a decrease in the percent of calories from total and saturated fats and an increase in percent of calories from carbohydrates, while there has been a relatively stable contribution of percent of calories from dietary protein (Wright et al. 2004).

As opposed to adults, self- and parent-reported mean energy intake for children aged 1–19 years changed little between 1971–1974 and 1999–2000, except for adolescent females (Briefel and Johnson 2004). For example, among boys 16–19 years, mean energy intake was 3,010 kcal in 1971–1974 and 2,932 kcal in 1999–2000 while among girls 16–19 years, mean energy intake was 1,735 kcal in 1971–1974 and 1,996 kcal in 1999–2000. In more recent analyses, comparing 1988–1994 and 2003–2006, mean caloric intake of adolescent females did not further increase (Lamb et al. 2009).

55.6.2 Global Trends in Nutrition

Similar to the United States, there was an increase (17%) in total food energy availability in Europe between 1961 and 2001 (Schmidhuber and Traill 2006). However, there has also been marked shifts in the structure of diets around the world, including South Korea (Kim et al. 2000), Brazil (Monteiro et al. 1995), China (Popkin 2001a), and India (Popkin et al. 2001). These dietary changes, sometimes referred to as the nutrition transition, are often unique to specific countries but may include increases in the consumption of fat and added sugar and decreases in dietary fiber (Popkin 2001b). Changing nutrition patterns and increasing obesity prevalence pose particularly unique challenges where obesity and undernutrition coexist, including Uzbekistan, Kiribati, and Algeria (de Onis and Blossner 2000).

55.6.3 Influence of Diet on Obesity

Among adults, increases in mean energy intake between 1971 and 2000 may partially explain the increase in obesity prevalence. However, no similar changes were seen among children. Additionally, further evidence suggests an increase in calories *does not* seem to explain the development of obesity among children (Eck et al. 1992; Troiano et al. 2000). One possible explanation for the lack of association between caloric intake and development of obesity is the limitation of the data. For example, much of the data for caloric intake is derived from 24-hour recalls and is subject to recall bias. Consumers' knowledge of the content of food and the ability to recall dietary intake may have changed since the 1970s, thereby limiting comparisons in dietary recall over time. Another possible explanation for the development of obesity may be the quality of one's diet including eating patterns and food composition, in addition to the quantity.

55.6.4 Eating Patterns

Between the 1970s and 1990s, there were increases in the consumption of meals away from home among Americans of all age (Nielsen et al. 2002). Meals consumed outside the home are associated with more calories (Guthrie et al. 2002) and higher dietary fat intake (Kearney et al. 2001; O'Dwyer et al. 2005) than meals consumed at home. One prospective cohort study conducted in Spain found a positive association between eating meals outside the home and the development of overweight and obesity (Bes-Rastrollo et al. 2009).

Between the 1970s and 1990s portion sizes increased in the United States for meals consumed both inside and outside the home (Nestle 2003; Nielsen and Popkin 2003). Larger portion sizes are associated with increased calories consumed during a meal (Diliberti et al. 2004) without compensatory decreases in calories consumed at subsequent meals (Kral and Rolls 2004; Rolls et al. 2004).

55.6.5 Composition of the Diet

55.6.5.1 Breast Milk

Breastfeeding for the first 4–6 months may protect individuals later in life against developing obesity (Institute of Medicine 2005; Owen et al. 2005). It has also been suggested that a longer duration of breastfeeding may further protect against developing obesity (Harder et al. 2005). There are several proposed mechanisms of protection against obesity including better ability to behaviorally respond to satiety cues (Fisher et al. 2000; Dewey 2003) and/or through biological activity of breast milk (Weyermann et al. 2007). In the absence of contraindications to breastfeeding, supporting breastfeeding has important benefits including its role in obesity prevention.

55.6.5.2 Dietary Fiber

Foods high in dietary fiber include fruits, vegetables, whole grains, legumes, and nuts. In cross-sectional studies, intake of dietary fiber is inversely associated with body weight (Miller et al. 1994; Nelson and Tucker 1996). In three prospective studies following participants between 6 and 14 years, diets with higher fiber content have been inversely associated with subsequent weight gain (Ludwig et al. 1999; Liu et al. 2003; Du et al. 2010). Research of specific dietary fiber sources have shown fruits and vegetables (He et al. 2004; Tang et al. 2010), whole grains (Koh-Banerjee et al. 2004), and nuts (Bes-Rastrollo et al. 2007) to also have an inverse association with obesity or weight gain (Hu 2008).

55.6.5.3 Sugar-Sweetened Beverages

In the United States, intake of added sugar accounts for 15.8% of total energy consumed, and the largest source of added sugars is from sugar-sweetened beverages, including carbonated sodas, fruit punch, and sports drinks (Guthrie and Morton 2000; Wang et al. 2008). A systemic review found that in feeding trials of adults, the intake of sugar-sweetened beverages supports weight gain (Malik et al. 2006). In prospective studies, sugar-sweetened beverages were associated with weight gain and obesity in children (Ludwig et al. 2001; Berkey et al. 2004) and adults (Schulze et al. 2004).

55.7 Contributing Factors to Obesity: Physical Activity

55.7.1 Temporal and Age-Related Trends in Physical Activity

In the United States, between the years 1950 and 2000, there were increases in the percent of adults who work in low physical activity occupations and declines in the proportions of persons who walk to work (Brownson et al. 2005). During the same time period, there were increases in the average daily vehicle miles traveled (Brownson et al. 2005). Changes in commuting patterns are not isolated to adults. Between 1969 and 2001, there were declines in the proportion of children who walked or rode a bicycle to school in the United States (McDonald 2007).

Data for national trends in leisure time physical activity in the United States are available from 1988 onward. Much of the physical activity, estimates are based on self-reported physical activity and individuals are characterized as physically inactive, i.e., no reported leisure time physical activity in the preceding month, or meeting recommended physical activity guidelines (United States Department of Health and Human Services 1996, 2008) and/or national goals (United States Department of Health and Human Services 2000). The data for physical inactivity are more comparable across years because of uniformity in the definition of physical inactivity. Trends in physical inactivity in the United States show a decline from 30.5% in 1988 to 25.1% in 2008 (CDC 2010) despite an increase in obesity prevalence.

Based on self-reported data, older Americans are more inactive than younger Americans (Kruger et al. 2005). Results of accelerometer-measured physical activity add to this finding. For example, children spend more time in physical activity than adolescents, and adolescents spend more time in physical activity than adults (Nader et al. 2008; Troiano et al. 2008). Further, accelerometer studies have shown that the age-related decline in physical activity largely occurs between 13 and 18 years and this decline is greater for men than women (Sallis 2000). Despite steeper declines in physical activity with aging for men, women have overall lower participation in physical activity in all age categories (i.e., childhood, adolescence, and adulthood) (Troiano et al. 2008).

55.7.2 Global Trends in Physical Activity

Similar to the United States, decreases in physical activity related to occupation and commute have been observed in other countries including European countries (Stamatakis et al. 2007; Lindahl et al. 2003), New Zealand (Tin Tin et al. 2009), Australia (Bell et al. 2006), and China (Popkin 2008). Increases in leisure time physical activity have been noted in Europe (Simmons et al. 1996; Lindstrom et al. 2003).

55.7.3 Association of Physical Activity with Obesity

Ecological studies comparing physical activity levels in multiple countries reveal an inverse association between physical activity and obesity (Bassett et al. 2008). Moreover, data from cross-sectional and longitudinal studies also demonstrate an inverse association between physical activity and both obesity and weight gain (Seidell et al. 1991; Lahti-Koski et al. 2002; Chan et al. 2003; Bernstein et al. 2004; United States Department of Health and Human Services 2008). However, the strength of these associations is tempered by the limited longitudinal data using direct measures of physical activity (Hu 2008). Objective measures of physical activity are preferable because persons tend to overreport physical activity levels (Troiano et al. 2008).

55.7.4 Sedentary Behaviors

Sedentary behaviors, including television viewing and computer/video game use, may impact health outcomes, including obesity (Dietz and Gortmaker 1985; Tucker and Friedman 1989; Tucker and Bagwell 1991; Ching et al. 1996; Jakes et al. 2003; Stettler et al. 2004), cardiovascular disease (Fung et al. 2000), and diabetes (Hu et al. 2003) independent of level of physical activity. Results from birth cohorts from New Zealand and Great Britain have found an association between early childhood television viewing and obesity in adulthood (Hancox et al. 2004; Viner and Cole 2005). Proposed mechanisms for the association between television viewing and obesity include increased snacking between meals and increased exposure to the effects of commercials (Thomson et al. 2008; Parvanta et al. 2010).

55.8 Contributing Factors to Obesity: Community and Home Environment

55.8.1 Community Factors

There is growing literature evaluating the role of community and home environments in obesity. One area of research that has focused on the community environment specifically examines the role of the "built environment." In the context of obesity, people often refer to the "built environment" as "access to grocery stores selling healthy foods, proximity and safety of playgrounds, and adequate housing" (Singh et al. 2010). Other community factors may impact dietary intake and physical activity levels among individuals. For example, access to affordable (French et al. 1997, 2001) and fresh food can impact food choices. Likewise, recreational facilities and community infrastructure to support bicycling, walking (Owen et al. 2004), and use of public transportation (Besser and Dannenberg 2005) can impact physical activity levels.

A walkability index has been used to research the relationship between the built environment and physical activity. One example of a commonly used walkability index is the Neighborhood Environment Walkability Scale (NEWS) (Saelens et al. 2003), a self-reported survey which assesses the following elements: residential density; proximity to, and ease of access to, non-residential land uses, such as restaurants and retail stores (land use mix-diversity and land use mix-access); street connectivity; walking/cycling facilities, such as sidewalks and pedestrian/bike trails; aesthetics; traffic safety; and crime safety. Other researchers have developed walkability indexes, based on geographic information systems (GIS), which have incorporated diverse measures including residential density, intersection density, and land use mix (Frank et al. 2010). Much of the research suggests that individuals living in a neighborhood with a high walkability index, defined by either NEWS (Saelens et al. 2003) or a GIS-derived scale (Frank et al. 2010; Van Dyck et al. 2010), are more active than individuals living in a neighborhood with a low walkability index.

55.8.2 Home Environment Factors

The association between parental and childhood obesity has been well described and can be attributed in part to a shared home environment (Strauss and Knight 1999; Kral and Rauh 2010) in addition to shared genetic factors (Faith et al. 1999). The home environment provides children an opportunity to model healthy lifestyle behaviors, including food consumption routines (Birch and Davison 2001) and dietary preferences (Fisher et al. 2002). A study examining the eating habits of preschool children showed that the parents' nutrient intake influenced those of their children (Oliveria et al. 1992). Another study, which included children ages 6–19 years, found similar associations (Laskarzewski et al. 1980).

A common feature of today's home environment is the television. American children on average view 1.8–2.8 h of television per day (Marshall et al. 2006). As a result, there is interest in exploring the components of television viewing, including television advertisements, and possible associations with obesity. In the United States (Powell et al. 2007), Mexico (Ramirez-Ley et al. 2009), Bulgaria (Galcheva et al. 2008), and elsewhere, food and beverage commercials are common. There is growing evidence that exposure to food and beverage advertisements may contribute to obesity in childhood and this influence cannot be entirely explained by reduced physical activity (Jackson et al. 2009; Zimmerman and Bell 2010).

55.9 Health Consequences of Obesity

Obesity has been linked to a wide range of health outcomes including hypertension, hypercholesterolemia, diabetes, gallbladder disease, osteoarthritis, sleep apnea, and cancers (National Institutes of Health 1998). Hypertension and hypercholesterolemia are important risk factors for the development of cardiovascular disease

(Clarke et al. 2009). Increased mortality from diabetes (Rogers et al. 2003), cancer (Flegal et al. 2007), and cardiovascular disease (Flegal et al. 2007) also has been attributed to obesity. In this section, mortality associated with overweight and obesity as well as select health consequences of obesity will be discussed. Further, the tracking of childhood obesity into adulthood will be highlighted.

55.9.1 Mortality

Compared with normal weight adults, obese adults have an increased all-cause mortality (Flegal et al. 2005; Berrington de Gonzalez et al. 2010). In a study of a representative sample of the United States population, this association was largely driven by the increased mortality from cardiovascular disease (CVD) among obese adults and, to a lesser extent, diabetes and kidney diseases combined, as well as cancers that have been epidemiologically associated with obesity which included colon, breast, esophageal, uterine, ovarian, kidney, and pancreatic cancers (Flegal et al. 2007). However, mortality from other cancers and non-cancer, non-CVDs were not associated with obesity. In the same study, Flegal et al. reported overweight, compared with normal weight, adults had a decreased all-cause mortality. Analysis of cause-specific mortality showed that overweight was not associated with mortality from cardiovascular disease or cancer and was associated with increased mortality from diabetes and kidney diseases combined and decreased non-cancer, non-CVD mortality. Other studies have reported increased mortality among overweight adults (Berrington de Gonzalez et al. 2010). However, differences in exclusion criteria and reference categories limit the ability to compare results across studies.

55.9.2 Hypertension

Hypertension, often defined as sustained systolic blood pressure of 140 mm of mercury (Hg) or greater or a diastolic blood pressure of 90 mm Hg or greater (Whitworth 2003; Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure 2004), is a risk factor for the development of a range of medical complications including coronary heart disease, cerebrovascular disease, renal disease, and congestive heart failure. Obesity is associated with hypertension in adults (MacMahon et al. 1987; Brown et al. 2000; Kotsis et al. 2010) and children (Kollias et al. 2009; Ostchega et al. 2009). Studies using a prospective cohort design have demonstrated increases in systolic and diastolic blood pressure with increases in BMI (Franklin et al. 2005). Further, clinical trials have shown that weight loss can lower blood pressure (Satterfield et al. 1991; The Trials of Hypertension Prevention Collaborative Research Group 1992, 1997). Worldwide, high blood pressure is responsible for 7.1 million premature deaths per year (World Health Organization 2002).

55.9.3 Hypercholesterolemia

By contributing to the development of arthrosclerosis (Stamler et al. 1986; Wilson et al. 1998), an elevation in serum cholesterol is associated with the development of coronary vascular disease and cerebrovascular disease. High total cholesterol is defined as a value greater than 240 mg/dL (National Cholesterol Education Program 2002). Overweight and obese (BMI $\geq$ 25 kg/m^2) persons are more likely to have an elevation in total cholesterol when compared with those with a healthy weight (Brown et al. 2000). Cholesterol levels track from childhood into adulthood (Lauer et al. 1988), and childhood cholesterol levels may predict the development of arthrosclerosis in adulthood (Li et al. 2003; Raitakari et al. 2003). Hypercholesterolemia has been estimated to be responsible for four million premature deaths per year (World Health Organization 2002).

55.9.4 Diabetes

Prevalence of type II diabetes, a state of hyperglycemia (elevated blood sugar) and relative insulin resistance, is higher among those with higher weight (Hartz et al. 1983; Colditz et al. 1995). In fact, obesity is a leading risk factor for the development of diabetes (National Institutes of Health 1998), and similar to obesity, there was an increased trend in diagnosed diabetes prevalence between 1970 and 1990 (Fox et al. 2006). Diabetes is a risk factor for the development of cardiovascular disease, end-stage renal disease, retinopathy, neuropathy, and other adverse health outcomes. Weight reduction, along with dietary and physical activity modifications and medications, is an important component of guideline-recommended care for diabetes treatment (American Diabetes Association 2010). Worldwide, a BMI above 21 kg/m^2 is estimated to account for approximately 58% of diabetes cases (World Health Organization 2002).

55.9.5 Cancer

Obesity has been associated with increased risk of cancers of the uterus, kidney, gallbladder, breast (postmenopausal women), esophagus, and colon (Calle et al. 2003). Obese individuals have a higher mortality from cancer types known to be epidemiologically associated with obesity (Flegal et al. 2007). Biological mechanisms for cancer risks are not fully understood and are likely different for different cancer types. However, some evidence suggests that there may be obesity-induced hormonal changes (World Health Organization 2002), including sex steroids, insulin, and insulin-like growth factors (Renehan et al. 2008). An estimated 2% of cancer mortality worldwide is associated with overweight and obesity (Danaei et al. 2005).

55.9.6 Tracking of Obesity from Childhood to Adulthood

Weight tracks from childhood into adulthood (Serdula et al. 1993; Fuentes et al. 2003). This means that obese children often become obese adults. Children as young as 2–5 years with an elevated BMI may be at higher risk for obesity during adulthood (Freedman et al. 2005). Further, obese children may have higher than expected levels of morbidity and mortality in adulthood (Mossberg 1989; Must et al. 1992; Gunnell et al. 1998). Given the propensity for young obese children to remain obese in adulthood, there is increased interest in understanding the influences of early childhood experiences in determining obesity outcomes (Reilly et al. 2005; Taveras et al. 2010).

55.10 Socioeconomic Consequences of Obesity

Overweight and obesity have a significant impact on individual economic and social achievements. Additionally, the collective individual health and socioeconomic consequences of obesity have important societal costs, including added health-care costs. While not individually discussed, there are other measures of increased societal cost including reduced work potential as a result of occupational absenteeism (Cawley et al. 2007) and excess morbidity and mortality.

55.10.1 Individuals

Lower achievement of advanced education (Wardle et al. 2002; Karnehed et al. 2006; Hajian-Tilaki and Heidari 2010), rates of employment (Klarenbach et al. 2006; Tunceli et al. 2006), and income (Baum and Ford 2004; Brunello and D'Hombres 2007; Staub et al. 2010) among obese individuals have been described. For example, one study showed that an increase in BMI of 10% was associated with a 3.3% and 1.9% lower hourly wage for men and women, respectively (Brunello and D'Hombres 2007). There are likely multiple mechanisms by which obesity and socioeconomic status are related. However, there has been a suggestion that negative attitudes, stigmatization, and stereotypes toward obese individuals are contributors to these differences (Puhl et al. 2008; Puhl and Heuer 2009).

In addition to differences in educational and economic achievements between obese and non-obese individuals, there are differences in other important social measures, including interpersonal relationships. Among children and adolescents, obesity is associated with lower level of satisfaction with life (Erickson et al. 2000; Franklin et al. 2006; BeLue et al. 2009) and increased difficulty making friends (Strauss and Pollack 2003; Fonseca et al. 2009). Among adults, reduced quality of relationships with friends (Ball et al. 2004) and romantic partners have been reported (Boyes and Latner 2009). Further, obese adults report lower health-related quality of life (Jia and Lubetkin 2005; Sach et al. 2007; Soltoft et al. 2009; Renzaho et al. 2010).

55.10.2 Health-Care System

Increased medical expenditures have been attributed to obesity. Higher use of primary care and other outpatient services (van Dijk et al. 2006), inpatient services (Finkelstein et al. 2009), and prescription medications (van Dijk et al. 2006; Finkelstein et al. 2009; Milder et al. 2010) among obese individuals contribute to the increased medical costs (Wolf and Colditz 1998). Estimates for the percentage of national health expenditures attributable to obesity differ by country. North American, Australian, and European data from the late 1980s and 1990s showed 2–6% of national health expenditures were attributed to obesity (Segal et al. 1994; Levy et al. 1995; Wolf and Colditz 1998; Birmingham et al. 1999). However, more recent data (2006) from both the United States and Canada show increases in national health expenditures attributable to obesity (Anis et al. 2010; Finkelstein et al. 2009).

55.11 Emerging Research

55.11.1 Prenatal and Early Infancy

Prenatal and early infancy factors may influence the development of obesity. For example, higher prevalence of obesity among children exposed in utero to excessive gestational weight gain (Oken et al. 2008; Wrotniak et al. 2008) and diabetes (Gillman et al. 2003; Wright et al. 2009) has been described. Birth weight may also be associated with obesity. However, the relationship between birth weight and obesity is complex and has not been consistently described in the literature. Some studies have found little or no association between birth weight and obesity (Parsons et al. 2001; The et al. 2010), while other research suggests both low birth weight and high birth weight are associated with obesity (Curhan et al. 1996a, b; Sorensen et al. 1997). Factors in early infancy may also be associated with obesity. These factors include duration of breastfeeding, which was previously described in this chapter, timing of introduction of solid foods (Baker et al. 2004; Seach et al. 2010), and rapidity of weight gain (Baird et al. 2005). More research into the prenatal and early infancy experiences is needed.

55.11.2 Sleep

Physiologic responses to sleep deprivation include elevated blood pressure (Tochikubo et al. 1996), impaired glucose tolerance (Spiegel et al. 1999), increased fatigue (Dinges et al. 1997), and increased hunger (Spiegel et al. 2004). Sleep deprivation may also predispose individuals to obesity (Patel and Hu 2008) and growing evidence suggests that shorter sleep duration may be associated with higher obesity prevalence in children (Gupta et al. 2002) and adults

(Gangwisch et al. 2005; Patel and Hu 2008). Further research, including more studies with objective measures of sleep duration, is warranted (Patel and Hu 2008).

55.12 Conclusions

Obesity prevalence has increased over the last 3 decades although the increase appears to have slowed in the United States and elsewhere. The magnitude of obesity is defined not only by its global prevalence but also by its health and socioeconomic consequences. Ongoing research to more fully characterize the dietary and physical activity patterns of individuals and societies continues to provide additional understanding into the factors that contribute to obesity. Knowledge of these factors informs interventions, including the United States' Surgeon General Vision for a Healthy and Fit Nation, 2010 (Benjamin 2010), to achieve maximal benefits.

The findings and conclusions in this report are those of the authors and not necessarily those of the Centers for Disease Control and Prevention.

References

Abraham S, Luscombe G, Boyd C, Olesen I (2004) Predictors of the accuracy of self-reported height and weight in adolescent female school students. Int J Eat Disord 36:76–82

Akinbami LJ, Ogden CL (2009) Childhood overweight prevalence in the United States: the impact of parent-reported height and weight. Obesity (Silver Spring) 17:1574–1580

American Diabetes Association (2010) Summary of revisions for the 2010 Clinical Practice Recommendations. Diabetes Care 33(Suppl 1):S3

Anderson DE (2007) Reliability of air displacement plethysmography. J Strength Cond Res 21:169–172

Anis AH, Zhang W, Bansback N, Guh DP, Amarsi Z, Birmingham CL (2010) Obesity and overweight in Canada: an updated cost-of-illness study. Obes Rev 11:31–40

Baird J, Fisher D, Lucas P, Kleijnen J, Roberts H, Law C (2005) Being big or growing fast: systematic review of size and growth in infancy and later obesity. BMJ 331:929

Baker JL, Michaelsen KF, Rasmussen KM, Sorensen TI (2004) Maternal prepregnant body mass index, duration of breastfeeding, and timing of complementary food introduction are associated with infant weight gain. Am J Clin Nutr 80:1579–1588

Balanza R, Garcia-Lorda P, Perez-Rodrigo C, Aranceta J, Bonet MB, Salas-Salvado J (2007) Trends in food availability determined by the Food and Agriculture Organization's food balance sheets in Mediterranean Europe in comparison with other European areas. Public Health Nutr 10:168–176

Ball K, Crawford D, Kenardy J (2004) Longitudinal relationships among overweight, life satisfaction, and aspirations in young women. Obes Res 12:1019–1030

Barlow SE, Dietz WH (1998) Obesity evaluation and treatment: expert Committee recommendations. The Maternal and Child Health Bureau, Health Resources and Services Administration and the Department of Health and Human Services. Pediatrics 102:E29

Bassett DR Jr, Pucher J, Buehler R, Thompson DL, Crouter SE (2008) Walking, cycling, and obesity rates in Europe, North America, and Australia. J Phys Act Health 5:795–814

Baum CL 2nd, Ford WF (2004) The wage effects of obesity: a longitudinal study. Health Econ 13:885–899

Bell AC, Garrard J, Swinburn BA (2006) Active transport to work in Australia: is it all downhill from here? Asia Pac J Public Health 18:62–68

BeLue R, Francis LA, Colaco B (2009) Mental health problems and overweight in a nationally representative sample of adolescents: effects of race and ethnicity. Pediatrics 123:697–702

Benjamin RM (2010) The Surgeon General's vision for a healthy and fit nation. Public Health Rep 125:514–515

Berkey CS, Rockett HR, Field AE, Gillman MW, Colditz GA (2004) Sugar-added beverages and adolescent weight change. Obes Res 12:778–788

Bernstein MS, Costanza MC, Morabia A (2004) Association of physical activity intensity levels with overweight and obesity in a population-based sample of adults. Prev Med 38:94–104

Berrington de Gonzalez A, Hartge P, Cerhan JR, Flint AJ, Hannan L, MacInnis RJ, Moore SC, Tobias GS, Anton-Culver H, Freeman LB, Beeson WL, Clipp SL, English DR, Folsom AR, Freedman DM, Giles G, Hakansson N, Henderson KD, Hoffman-Bolton J, Hoppin JA, Koenig KL, Lee IM, Linet MS, Park Y, Pocobelli G, Schatzkin A, Sesso HD, Weiderpass E, Willcox BJ, Wolk A, Zeleniuch-Jacquotte A, Willett WC, Thun MJ (2010) Body-mass index and mortality among 1.46 million white adults. N Engl J Med 363:2211–2219

Bes-Rastrollo M, Sabate J, Gomez-Gracia E, Alonso A, Martinez JA, Martinez-Gonzalez MA (2007) Nut consumption and weight gain in a Mediterranean cohort: the SUN study. Obesity (Silver Spring) 15:107–116

Bes-Rastrollo M, Basterra-Gortari FJ, Sanchez-Villegas A, Marti A, Martinez JA, Martinez-Gonzalez MA (2009) A prospective study of eating away-from-home meals and weight gain in a Mediterranean population: the SUN (Seguimiento Universidad de Navarra) cohort. Public Health Nutr 13:1356–1363

Besser LM, Dannenberg AL (2005) Walking to public transit: steps to help meet physical activity recommendations. Am J Prev Med 29:273–280

Birch LL, Davison KK (2001) Family environmental factors influencing the developing behavioral controls of food intake and childhood overweight. Pediatr Clin North Am 48:893–907

Birmingham CL, Muller JL, Palepu A, Spinelli JJ, Anis AH (1999) The cost of obesity in Canada. CMAJ 160:483–488

Bostrom G, Diderichsen F (1997) Socioeconomic differentials in misclassification of height, weight and body mass index based on questionnaire data. Int J Epidemiol 26:860–866

Boyes AD, Latner JD (2009) Weight stigma in existing romantic relationships. J Sex Marital Ther 35:282–293

Bray GA, Greenway FL, Molitch ME, Dahms WT, Atkinson RL, Hamilton K (1978) Use of anthropometric measures to assess weight loss. Am J Clin Nutr 31:769–773

Brenner DJ, Hall EJ (2007) Computed tomography–an increasing source of radiation exposure. N Engl J Med 357:2277–2284

Briefel RR, Johnson CL (2004) Secular trends in dietary intake in the United States. Annu Rev Nutr 24:401–431

Brown CD, Higgins M, Donato KA, Rohde FC, Garrison R, Obarzanek E, Ernst ND, Horan M (2000) Body mass index and the prevalence of hypertension and dyslipidemia. Obes Res 8: 605–619

Brownson RC, Boehmer TK, Luke DA (2005) Declining rates of physical activity in the United States: what are the contributors? Annu Rev Public Health 26:421–443

Brozek J, Grande F, Anderson JT, Keys A (1963) Densitometric analysis of body composition: revision of some quantitative assumptions. Ann N Y Acad Sci 110:113–140

Brunello G, D'Hombres B (2007) Does body weight affect wages? Evidence from Europe. Econ Hum Biol 5:1–19

Cacciari E, Milani S, Balsamo A, Spada E, Bona G, Cavallo L, Cerutti F, Gargantini L, Greggio N, Tonini G, Cicognani A (2006) Italian cross-sectional growth charts for height, weight and BMI (2 to 20 yr). J Endocrinol Invest 29:581–593

Calle EE, Rodriguez C, Walker-Thurmond K, Thun MJ (2003) Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. N Engl J Med 348:1625–1638

Cawley J, Rizzo JA, Haas K (2007) Occupation-specific absenteeism costs associated with obesity and morbid obesity. J Occup Environ Med 49:1317–1324

CDC (2010) Physical activity statistics: no leisure time physical activity trend chart 2010. http://www.cdc.gov/nccdphp/dnpa/physical/stats/leisure_time.htm. Accessed 10 June 2010

Chan CB, Spangler E, Valcour J, Tudor-Locke C (2003) Cross-sectional relationship of pedometer-determined ambulatory activity to indicators of health. Obes Res 11:1563–1570

Ching PL, Willett WC, Rimm EB, Colditz GA, Gortmaker SL, Stampfer MJ (1996) Activity level and risk of overweight in male health professionals. Am J Public Health 86:25–30

Clarke R, Emberson J, Fletcher A, Breeze E, Marmot M, Shipley MJ (2009) Life expectancy in relation to cardiovascular risk factors: 38 year follow-up of 19,000 men in the Whitehall study. BMJ 339:b3513

Colditz GA, Willett WC, Rotnitzky A, Manson JE (1995) Weight gain as a risk factor for clinical diabetes mellitus in women. Ann Intern Med 122:481–486

Cole TJ, Freeman JV, Preece MA (1998) British 1990 growth reference centiles for weight, height, body mass index and head circumference fitted by maximum penalized likelihood. Stat Med 17:407–429

Cole TJ, Bellizzi MC, Flegal KM, Dietz WH (2000) Establishing a standard definition for child overweight and obesity worldwide: international survey. BMJ 320:1240–1243

Cordero-MacIntyre ZR, Peters W, Libanati CR, Espana RC, Abila SO, Howell WH, Lohman TG (2002) Reproducibility of DXA in obese women. J Clin Densitom 5:35–44

Covey MK, Berry JK, Hacker ED (2010) Regional body composition: cross-calibration of DXA scanners–QDR4500W and Discovery Wi. Obesity (Silver Spring) 18:632–637

Curhan GC, Chertow GM, Willett WC, Spiegelman D, Colditz GA, Manson JE, Speizer FE, Stampfer MJ (1996a) Birth weight and adult hypertension and obesity in women. Circulation 94:1310–1315

Curhan GC, Willett WC, Rimm EB, Spiegelman D, Ascherio AL, Stampfer MJ (1996b) Birth weight and adult hypertension, diabetes mellitus, and obesity in US men. Circulation 94: 3246–3250

Danaei G, Vander Hoorn S, Lopez AD, Murray CJ, Ezzati M (2005) Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. Lancet 366:1784–1793

de Onis M, Blossner M (2000) Prevalence and trends of overweight among preschool children in developing countries. Am J Clin Nutr 72:1032–1039

de Onis M, Blossner M (2003) The World Health Organization Global Database on Child Growth and Malnutrition: methodology and applications. Int J Epidemiol 32:518–526

de Onis M, Garza C, Victora CG, Onyango AW, Frongillo EA, Martines J (2004) The WHO Multicentre Growth Reference Study: planning, study design, and methodology. Food Nutr Bull 25:S15–26

de Onis M, Onyango AW, Borghi E, Siyam A, Nishida C, Siekmann J (2007) Development of a WHO growth reference for school-aged children and adolescents. Bull World Health Organ 85:660–667

Deurenberg P, Deurenberg-Yap M, Guricci S (2002) Asians are different from Caucasians and from each other in their body mass index/body fat per cent relationship. Obes Rev 3:141–146

Dewey KG (2003) Is breastfeeding protective against child obesity? J Hum Lact 19:9–18

Dietz WH Jr, Gortmaker SL (1985) Do we fatten our children at the television set? Obesity and television viewing in children and adolescents. Pediatrics 75:807–812

Diliberti N, Bordi PL, Conklin MT, Roe LS, Rolls BJ (2004) Increased portion size leads to increased energy intake in a restaurant meal. Obes Res 12:562–568

Dinges DF, Pack F, Williams K, Gillen KA, Powell JW, Ott GE, Aptowicz C, Pack AI (1997) Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4–5 hours per night. Sleep 20:267–277

Du H, van der AD, Boshuizen HC, Forouhi NG, Wareham NJ, Halkjaer J, Tjonneland A, Overvad K, Jakobsen MU, Boeing H, Buijsse B, Masala G, Palli D, Sorensen TI, Saris WH, Feskens EJ (2010) Dietary fiber and subsequent changes in body weight and waist circumference in European men and women. Am J Clin Nutr 91:329–336

Eck LH, Klesges RC, Hanson CL, Slawson D (1992) Children at familial risk for obesity: an examination of dietary intake, physical activity and weight status. Int J Obes Relat Metab Disord 16:71–78

Erickson SJ, Robinson TN, Haydel KF, Killen JD (2000) Are overweight children unhappy?: Body mass index, depressive symptoms, and overweight concerns in elementary school children. Arch Pediatr Adolesc Med 154:931–935

Evans WJ, Campbell WW (1993) Sarcopenia and age-related changes in body composition and functional capacity. J Nutr 123:465–468

Faith MS, Pietrobelli A, Nunez C, Heo M, Heymsfield SB, Allison DB (1999) Evidence for independent genetic influences on fat mass and body mass index in a pediatric twin sample. Pediatrics 104:61–67

Fakhrawi DH, Beeson L, Libanati C, Feleke D, Kim H, Quansah A, Darnell A, Lammi-Keefe CJ, Cordero-MacIntyre Z (2009) Comparison of body composition by bioelectrical impedance and dual-energy x-ray absorptiometry in overweight/obese postmenopausal women. J Clin Densitom 12:238–244

Fields DA, Hunter GR, Goran MI (2000) Validation of the BOD POD with hydrostatic weighing: influence of body clothing. Int J Obes Relat Metab Disord 24:200–205

Finkelstein EA, Trogdon JG, Cohen JW, Dietz W (2009) Annual medical spending attributable to obesity: payer-and service-specific estimates. Health Aff (Millwood) 28:w822–831

Fisher JO, Birch LL, Smiciklas-Wright H, Picciano MF (2000) Breast-feeding through the first year predicts maternal control in feeding and subsequent toddler energy intakes. J Am Diet Assoc 100:641–646

Fisher JO, Mitchell DC, Smiciklas-Wright H, Birch LL (2002) Parental influences on young girls' fruit and vegetable, micronutrient, and fat intakes. J Am Diet Assoc 102:58–64

Flegal KM, Carroll MD, Ogden CL, Johnson CL (2002) Prevalence and trends in obesity among US adults, 1999–2000. JAMA 288:1723–1727

Flegal KM, Graubard BI, Williamson DF, Gail MH (2005) Excess deaths associated with underweight, overweight, and obesity. JAMA 293:1861–1867

Flegal KM, Graubard BI, Williamson DF, Gail MH (2007) Cause-specific excess deaths associated with underweight, overweight, and obesity. JAMA 298:2028–2037

Flegal KM, Shepherd JA, Looker AC, Graubard BI, Borrud LG, Ogden CL, Harris TB, Everhart JE, Schenker N (2009) Comparisons of percentage body fat, body mass index, waist circumference, and waist-stature ratio in adults. Am J Clin Nutr 89:500–508

Flegal KM, Carroll MD, Ogden CL, Curtin LR (2010a) Prevalence and trends in obesity among US adults, 1999–2008. JAMA 303:235–241

Flegal KM, Ogden CL, Yanovski JA, Freedman DS, Shepherd JA, Graubard BI, Borrud LG (2010b) High adiposity and high body mass index-for-age in US children and adolescents overall and by race-ethnic group. Am J Clin Nutr 91:1020–1026

Flegal KM, Ogden CL (2011) Childhood obesity: are we all speaking the same language? Adv Nutr 2:159S–166S.

Fonseca H, Matos MG, Guerra A, Pedro JG (2009) Are overweight and obese adolescents different from their peers? Int J Pediatr Obes 4:166–174

Fox CS, Pencina MJ, Meigs JB, Vasan RS, Levitzky YS, D'Agostino RB Sr (2006) Trends in the incidence of type 2 diabetes mellitus from the 1970s to the 1990s: the Framingham Heart Study. Circulation 113:2914–2918

Frank LD, Sallis JF, Saelens BE, Leary L, Cain K, Conway TL, Hess PM (2010) The development of a walkability index: application to the Neighborhood Quality of Life Study. Br J Sports Med 44:924–933

Franklin SS, Pio JR, Wong ND, Larson MG, Leip EP, Vasan RS, Levy D (2005) Predictors of new-onset diastolic and systolic hypertension: the Framingham Heart Study. Circulation 111: 1121–1127

Franklin J, Denyer G, Steinbeck KS, Caterson ID, Hill AJ (2006) Obesity and risk of low self-esteem: a statewide survey of Australian children. Pediatrics 118:2481–2487

Freedman DS, Khan LK, Serdula MK, Dietz WH, Srinivasan SR, Berenson GS (2005) The relation of childhood BMI to adult adiposity: the Bogalusa Heart Study. Pediatrics 115:22–27

French SA, Jeffery RW, Story M, Hannan P, Snyder MP (1997) A pricing strategy to promote low-fat snack choices through vending machines. Am J Public Health 87:849–851

French SA, Jeffery RW, Story M, Breitlow KK, Baxter JS, Hannan P, Snyder MP (2001) Pricing and promotion effects on low-fat vending snack purchases: the CHIPS Study. Am J Public Health 91:112–117

Fryar CD, Carroll M, Ogden CL (2012a) Prevalence of obesity among children and adolescents: United States, trends 1963–1965 through 2009–2010. http://www.cdc.gov/nchs/data/hestat/obesity_child_09_10/obesity_child_09_10.pdf. Accessed 30 May 2013

Fryar CD, Carroll M, Ogden CL (2012b) Prevalence of overweight, obesity, and extreme obesity among adults: United States, trends 1976–1980 through 2009–2010. http://www.cdc.gov/nchs/data/hestat/obesity_adult_09_10/obesity_adult_09_10.pdf. Accessed 30 June 2013

Fuentes RM, Notkola IL, Shemeikka S, Tuomilehto J, Nissinen A (2003) Tracking of body mass index during childhood: a 15-year prospective population-based family study in eastern Finland. Int J Obes Relat Metab Disord 27:716–721

Fung TT, Hu FB, Yu J, Chu NF, Spiegelman D, Tofler GH, Willett WC, Rimm EB (2000) Leisure-time physical activity, television watching, and plasma biomarkers of obesity and cardiovascular disease risk. Am J Epidemiol 152:1171–1178

Galcheva SV, Iotova VM, Stratev VK (2008) Television food advertising directed towards Bulgarian children. Arch Dis Child 93:857–861

Gallagher D, Visser M, Sepulveda D, Pierson RN, Harris T, Heymsfield SB (1996) How useful is body mass index for comparison of body fatness across age, sex, and ethnic groups? Am J Epidemiol 143:228–239

Gallagher D, Visser M, De Meersman RE, Sepulveda D, Baumgartner RN, Pierson RN, Harris T, Heymsfield SB (1997) Appendicular skeletal muscle mass: effects of age, gender, and ethnicity. J Appl Physiol 83:229–239

Gangwisch JE, Malaspina D, Boden-Albala B, Heymsfield SB (2005) Inadequate sleep as a risk factor for obesity: analyses of the NHANES I. Sleep 28:1289–1296

Gillman MW, Rifas-Shiman S, Berkey CS, Field AE, Colditz GA (2003) Maternal gestational diabetes, birth weight, and adolescent obesity. Pediatrics 111:e221–226

Gillum RF, Sempos CT (2005) Ethnic variation in validity of classification of overweight and obesity using self-reported weight and height in American women and men: the Third National Health and Nutrition Examination Survey. Nutr J 4:27

Gunnell DJ, Frankel SJ, Nanchahal K, Peters TJ, Davey Smith G (1998) Childhood obesity and adult cardiovascular mortality: a 57-y follow-up study based on the Boyd Orr cohort. Am J Clin Nutr 67:1111–1118

Gupta NK, Mueller WH, Chan W, Meininger JC (2002) Is obesity associated with poor sleep quality in adolescents? Am J Hum Biol 14:762–768

Guthrie JF, Morton JF (2000) Food sources of added sweeteners in the diets of Americans. J Am Diet Assoc 100:43–51, quiz 49–50

Guthrie JF, Lin BH, Frazao E (2002) Role of food prepared away from home in the American diet, 1977–78 versus 1994–96: changes and consequences. J Nutr Educ Behav 34:140–150

Hajian-Tilaki KO, Heidari B (2010) Association of educational level with risk of obesity and abdominal obesity in Iranian adults. J Public Health (Oxf) 32:202–209

Hancox RJ, Milne BJ, Poulton R (2004) Association between child and adolescent television viewing and adult health: a longitudinal birth cohort study. Lancet 364:257–262

Harder T, Bergmann R, Kallischnigg G, Plagemann A (2005) Duration of breastfeeding and risk of overweight: a meta-analysis. Am J Epidemiol 162:397–403

Harnack LJ, Jeffery RW, Boutelle KN (2000) Temporal trends in energy intake in the United States: an ecologic perspective. Am J Clin Nutr 71:1478–1484

Hartz AJ, Rupley DC Jr, Kalkhoff RD, Rimm AA (1983) Relationship of obesity to diabetes: influence of obesity level and body fat distribution. Prev Med 12:351–357

Haug E, Rasmussen M, Samdal O, Iannotti R, Kelly C, Borraccino A, Vereecken C, Melkevik O, Lazzeri G, Giacchi M, Ercan O, Due P, Ravens-Sieberer U, Currie C, Morgan A, Ahluwalia N (2009) Overweight in school-aged children and its relationship with demographic and lifestyle factors: results from the WHO-Collaborative Health Behaviour in School-aged Children (HBSC) study. Int J Public Health 54(Suppl 2):167–179

He K, Hu FB, Colditz GA, Manson JE, Willett WC, Liu S (2004) Changes in intake of fruits and vegetables in relation to risk of obesity and weight gain among middle-aged women. Int J Obes Relat Metab Disord 28:1569–1574

Himes JH, Dietz WH (1994) Guidelines for overweight in adolescent preventive services: recommendations from an expert committee. The Expert Committee on Clinical Guidelines for Overweight in Adolescent Preventive Services. Am J Clin Nutr 59:307–316

Hu FB, Li TY, Colditz GA, Willett WC, Manson JE (2003) Television watching and other sedentary behaviors in relation to risk of obesity and type 2 diabetes mellitus in women. JAMA 289: 1785–1791

Hu FB (2008) Obesity epidemiology Oxford University Press, New York.

Institute of Medicine (2005) Preventing childhood obesity: health in the balance. The National Academies Press, Washington, DC

Jackson AS, Pollock ML (1978) Generalized equations for predicting body density of men. Br J Nutr 40:497–504

Jackson DM, Djafarian K, Stewart J, Speakman JR (2009) Increased television viewing is associated with elevated body fatness but not with lower total energy expenditure in children. Am J Clin Nutr 89:1031–1036

Jakes RW, Day NE, Khaw KT, Luben R, Oakes S, Welch A, Bingham S, Wareham NJ (2003) Television viewing and low participation in vigorous recreation are independently associated with obesity and markers of cardiovascular disease risk: EPIC-Norfolk population-based study. Eur J Clin Nutr 57:1089–1096

Janssen I, Katzmarzyk PT, Ross R (2004) Waist circumference and not body mass index explains obesity-related health risk. Am J Clin Nutr 79:379–384

Jia H, Lubetkin EI (2005) The impact of obesity on health-related quality-of-life in the general adult US population. J Public Health (Oxf) 27:156–164

Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (2004) The seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure: complete report. US Department of Health and Human Services, Bethesda

Karnehed N, Rasmussen F, Hemmingsson T, Tynelius P (2006) Obesity and attained education: cohort study of more than 700,000 Swedish men. Obesity (Silver Spring) 14:1421–1428

Kearney JM, Hulshof KF, Gibney MJ (2001) Eating patterns–temporal distribution, converging and diverging foods, meals eaten inside and outside of the home–implications for developing FBDG. Public Health Nutr 4:693–698

Kiebzak GM, Leamy LJ, Pierson LM, Nord RH, Zhang ZY (2000) Measurement precision of body composition variables using the lunar DPX-L densitometer. J Clin Densitom 3:35–41

Kim S, Moon S, Popkin BM (2000) The nutrition transition in South Korea. Am J Clin Nutr 71: 44–53

Klarenbach S, Padwal R, Chuck A, Jacobs P (2006) Population-based analysis of obesity and workforce participation. Obesity (Silver Spring) 14:920–927

Koh-Banerjee P, Franz M, Sampson L, Liu S, Jacobs DR Jr, Spiegelman D, Willett W, Rimm E (2004) Changes in whole-grain, bran, and cereal fiber consumption in relation to 8-y weight gain among men. Am J Clin Nutr 80:1237–1245

Kollias A, Antonodimitrakis P, Grammatikos E, Chatziantonakis N, Grammatikos EE, Stergiou GS (2009) Trends in high blood pressure prevalence in Greek adolescents. J Hum Hypertens 23:385–390

Kotsis V, Stabouli S, Papakatsika S, Rizos Z, Parati G (2010) Mechanisms of obesity-induced hypertension. Hypertens Res 33:386–393

Kral TV, Rauh EM (2010) Eating behaviors of children in the context of their family environment. Physiol Behav 100:567–573

Kral TV, Rolls BJ (2004) Energy density and portion size: their independent and combined effects on energy intake. Physiol Behav 82:131–138

Krebs NF, Himes JH, Jacobson D, Nicklas TA, Guilday P, Styne D (2007) Assessment of child and adolescent overweight and obesity. Pediatrics 120(Suppl 4):S193–228

Kruger J, Ham SA, Kohl HW (2005) Trends in leisure-time physical inactivity by age, sex, and race/ethnicity–United States, 1994–2004. MMWR Morb Mortal Wkly Rep 54:991–994

Kuczmarski RJ, Ogden CL, Guo SS, Grummer-Strawn LM, Flegal KM, Mei Z, Wei R, Curtin LR, Roche AF, Johnson CL (2002) 2000 CDC Growth Charts for the United States: methods and development. Vital Health Stat 11:1–190

Lahti-Koski M, Pietinen P, Heliovaara M, Vartiainen E (2002) Associations of body mass index and obesity with physical activity, food choices, alcohol intake, and smoking in the 1982–1997 FINRISK Studies. Am J Clin Nutr 75:809–817

Lamb MM, Carroll MD, Ogden CL (2009) Overweight/obesity among school-aged youth in the United States. The State Education Standard 10:4–13. www.nasbe.org. Accessed 15 Jan 2011

Laskarzewski P, Morrison JA, Khoury P, Kelly K, Glatfelter L, Larsen R, Glueck CJ (1980) Parent-child nutrient intake interrelationships in school children ages 6 to 19: the Princeton School District Study. Am J Clin Nutr 33:2350–2355

Lauer RM, Lee J, Clarke WR (1988) Factors affecting the relationship between childhood and adult cholesterol levels: the Muscatine Study. Pediatrics 82:309–318

Levy E, Levy P, Le Pen C, Basdevant A (1995) The economic cost of obesity: the French situation. Int J Obes Relat Metab Disord 19:788–792

Li S, Chen W, Srinivasan SR, Bond MG, Tang R, Urbina EM, Berenson GS (2003) Childhood cardiovascular risk factors and carotid vascular changes in adulthood: the Bogalusa Heart Study. JAMA 290:2271–2276

Lindahl B, Stegmayr B, Johansson I, Weinehall L, Hallmans G (2003) Trends in lifestyle 1986–99 in a 25- to 64-year-old population of the Northern Sweden MONICA project. Scand J Public Health Suppl 61:31–37

Lindstrom M, Isacsson SO, Merlo J (2003) Increasing prevalence of overweight, obesity and physical inactivity: two population-based studies 1986 and 1994. Eur J Public Health 13: 306–312

Liu S, Willett WC, Manson JE, Hu FB, Rosner B, Colditz G (2003) Relation between changes in intakes of dietary fiber and grain products and changes in weight and development of obesity among middle-aged women. Am J Clin Nutr 78:920–927

Lohman TG (1981) Skinfolds and body density and their relation to body fatness: a review. Hum Biol 53:181–225

Ludwig DS, Pereira MA, Kroenke CH, Hilner JE, Van Horn L, Slattery ML, Jacobs DR Jr (1999) Dietary fiber, weight gain, and cardiovascular disease risk factors in young adults. JAMA 282:1539–1546

Ludwig DS, Peterson KE, Gortmaker SL (2001) Relation between consumption of sugar-sweetened drinks and childhood obesity: a prospective, observational analysis. Lancet 357:505–508

MacMahon S, Cutler J, Brittain E, Higgins M (1987) Obesity and hypertension: epidemiological and clinical issues. Eur Heart J 8(Suppl B):57–70

Malik VS, Schulze MB, Hu FB (2006) Intake of sugar-sweetened beverages and weight gain: a systematic review. Am J Clin Nutr 84:274–288

Marks GC, Habicht JP, Mueller WH (1989) Reliability, dependability, and precision of anthropometric measurements. The Second National Health and Nutrition Examination Survey 1976–1980. Am J Epidemiol 130:578–587

Marshall SJ, Gorely T, Biddle SJ (2006) A descriptive epidemiology of screen-based media use in youth: a review and critique. J Adolesc 29:333–349

McDonald NC (2007) Active transportation to school: trends among U.S. schoolchildren, 1969–2001. Am J Prev Med 32:509–516

Milder IE, Klungel OH, Mantel-Teeuwisse AK, Verschuren WM, Bemelmans WJ (2010) Relation between body mass index, physical inactivity and use of prescription drugs: the Doetinchem Cohort Study. Int J Obes (Lond) 34:1060–1069

Miller WC, Niederpruem MG, Wallace JP, Lindeman AK (1994) Dietary fat, sugar, and fiber predict body fat content. J Am Diet Assoc 94:612–615

Monteiro CA, Mondini L, de Souza AL, Popkin BM (1995) The nutrition transition in Brazil. Eur J Clin Nutr 49:105–113

Moreno LA, Pigeot I, Ahrens W (2011) Epidemiology of obesity in children and adolescents – prevalence and etiology. Springer, Berlin/Heidelberg/New York

Mossberg HO (1989) 40-year follow-up of overweight children. Lancet 2:491–493

Must A, Jacques PF, Dallal GE, Bajema CJ, Dietz WH (1992) Long-term morbidity and mortality of overweight adolescents. A follow-up of the Harvard Growth Study of 1922 to 1935. N Engl J Med 327:1350–1355

Nader PR, Bradley RH, Houts RM, McRitchie SL, O'Brien M (2008) Moderate-to-vigorous physical activity from ages 9 to 15 years. JAMA 300:295–305

National Cholesterol Education Program (USA) (2002) Third report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III): final report. National Institutes of Health, Washington, DC

National Institutes of Health (1998) Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults: executive summary. Expert Panel on the Identification, Evaluation, and Treatment of Overweight in Adults. Am J Clin Nutr 68: 899–917

Nelson LH, Tucker LA (1996) Diet composition related to body fat in a multivariate study of 203 men. J Am Diet Assoc 96:771–777

Nestle M (2003) Increasing portion sizes in American diets: more calories, more obesity. J Am Diet Assoc 103:39–40

Nielsen SJ, Popkin BM (2003) Patterns and trends in food portion sizes, 1977–1998. JAMA 289:450–453

Nielsen SJ, Siega-Riz AM, Popkin BM (2002) Trends in energy intake in U.S. between 1977 and 1996: similar shifts seen across age groups. Obes Res 10:370–378

Noreen EE, Lemon PW (2006) Reliability of air displacement plethysmography in a large, heterogeneous sample. Med Sci Sports Exerc 38:1505–1509

O'Dwyer NA, Gibney MJ, Burke SJ, McCarthy SN (2005) The influence of eating location on nutrient intakes in Irish adults: implications for developing food-based dietary guidelines. Public Health Nutr 8:258–265

Ogden CL, Carroll MD, Curtin LR, Lamb MM, Flegal KM (2010) Prevalence of high body mass index in US children and adolescents, 2007–2008. JAMA 303:242–249

Oken E, Rifas-Shiman SL, Field AE, Frazier AL, Gillman MW (2008) Maternal gestational weight gain and offspring weight in adolescence. Obstet Gynecol 112:999–1006

Olaiz-Fernàndez G R-DJ, Shamah-Levy T, Rojas R, Villalpando-Hernàndez S, Hernàndez-Avila M, Sepúlveda-Amor J (2006) Encuesta Nacional de Salud y Nutrición. http://www.todoendiabetes.org/diabe2/pdf/ensanut2006.pdf. Accessed 10 June 2010

Olds TS, Tomkinson GR, Ferrar KE, Maher CA (2010) Trends in the prevalence of childhood overweight and obesity in Australia between 1985 and 2008. Int J Obes (Lond) 34:57–66

Oliveria SA, Ellison RC, Moore LL, Gillman MW, Garrahie EJ, Singer MR (1992) Parent-child relationships in nutrient intake: the Framingham Children's Study. Am J Clin Nutr 56:593–598

Ostchega Y, Carroll M, Prineas RJ, McDowell MA, Louis T, Tilert T (2009) Trends of elevated blood pressure among children and adolescents: data from the National Health and Nutrition Examination Survey 1988–2006. Am J Hypertens 22:59–67

Owen N, Humpel N, Leslie E, Bauman A, Sallis JF (2004) Understanding environmental influences on walking; Review and research agenda. Am J Prev Med 27:67–76

Owen CG, Martin RM, Whincup PH, Smith GD, Cook DG (2005) Effect of infant feeding on the risk of obesity across the life course: a quantitative review of published evidence. Pediatrics 115:1367–1377

Parsons TJ, Power C, Manor O (2001) Fetal and early life growth and body mass index from birth to early adulthood in 1958 British cohort: longitudinal study. BMJ 323:1331–1335

Parvanta SA, Brown JD, Du S, Zimmer CR, Zhao X, Zhai F (2010) Television use and snacking behaviors among children and adolescents in China. J Adolesc Health 46:339–345

Patel SR, Hu FB (2008) Short sleep duration and weight gain: a systematic review. Obesity (Silver Spring) 16:643–653

Pateyjohns IR, Brinkworth GD, Buckley JD, Noakes M, Clifton PM (2006) Comparison of three bioelectrical impedance methods with DXA in overweight and obese men. Obesity (Silver Spring) 14:2064–2070

Peterson MJ, Czerwinski SA, Siervogel RM (2003) Development and validation of skinfold-thickness prediction equations with a 4-compartment model. Am J Clin Nutr 77:1186–1191

Plourde G (1997) The role of radiologic methods in assessing body composition and related metabolic parameters. Nutr Rev 55:289–296

Popkin BM (2001a) Nutrition in transition: the changing global nutrition challenge. Asia Pac J Clin Nutr 10(Suppl):S13–18

Popkin BM (2001b) The nutrition transition and obesity in the developing world. J Nutr 131: 871S–873S

Popkin BM (2008) Will China's nutrition transition overwhelm its health care system and slow economic growth? Health Aff (Millwood) 27:1064–1076

Popkin BM, Horton S, Kim S, Mahal A, Shuigao J (2001) Trends in diet, nutritional status, and diet-related noncommunicable diseases in China and India: the economic costs of the nutrition transition. Nutr Rev 59:379–390

Powell LM, Szczypka G, Chaloupka FJ, Braunschweig CL (2007) Nutritional content of television food advertisements seen by children and adolescents in the United States. Pediatrics 120: 576–583

Puhl RM, Heuer CA (2009) The stigma of obesity: a review and update. Obesity (Silver Spring) 17:941–964

Puhl RM, Andreyeva T, Brownell KD (2008) Perceptions of weight discrimination: prevalence and comparison to race and gender discrimination in America. Int J Obes (Lond) 32:992–1000

Putnam J, Allshouse J, Kantor LS (2002) U.S. per capita food supply trends: more calories, refined carbohydrates, and fats. Food Rev 25:2–15

Raitakari OT, Juonala M, Kahonen M, Taittonen L, Laitinen T, Maki-Torkko N, Jarvisalo MJ, Uhari M, Jokinen E, Ronnemaa T, Akerblom HK, Viikari JS (2003) Cardiovascular risk factors in childhood and carotid artery intima-media thickness in adulthood: the Cardiovascular Risk in Young Finns Study. JAMA 290:2277–2283

Ramirez-Ley K, De Lira-Garcia C, Souto-Gallardo Mde L, Tejeda-Lopez MF, Castaneda-Gonzalez LM, Bacardi-Gascon M, Jimenez-Cruz A (2009) Food-related advertising geared toward Mexican children. J Public Health (Oxf) 31:383–388

Reilly JJ, Armstrong J, Dorosty AR, Emmett PM, Ness A, Rogers I, Steer C, Sherriff A (2005) Early life risk factors for obesity in childhood: cohort study. BMJ 330:1357

Renehan AG, Roberts DL, Dive C (2008) Obesity and cancer: pathophysiological and biological mechanisms. Arch Physiol Biochem 114:71–83

Renzaho A, Wooden M, Houng B (2010) Associations between body mass index and health-related quality of life among Australian adults. Qual Life Res 19:515–520

Rogers RG, Hummer RA, Krueger PM (2003) The effect of obesity on overall, circulatory disease- and diabetes-specific mortality. J Biosoc Sci 35:107–129

Rolland-Cachera MF, Deheeger M, Bellisle F, Sempe M, Guilloud-Bataille M, Patois E (1984) Adiposity rebound in children: a simple indicator for predicting obesity. Am J Clin Nutr 39:129–135

Rolland-Cachera MF, Castetbon K, Arnault N, Bellisle F, Romano MC, Lehingue Y, Frelut ML, Hercberg S (2002) Body mass index in 7–9-y-old French children: frequency of obesity, overweight and thinness. Int J Obes Relat Metab Disord 26:1610–1616

Rolls BJ, Roe LS, Kral TV, Meengs JS, Wall DE (2004) Increasing the portion size of a packaged snack increases energy intake in men and women. Appetite 42:63–69

Sach TH, Barton GR, Doherty M, Muir KR, Jenkinson C, Avery AJ (2007) The relationship between body mass index and health-related quality of life: comparing the EQ-5D, EuroQol VAS and SF-6D. Int J Obes (Lond) 31:189–196

Saelens BE, Sallis JF, Black JB, Chen D (2003) Neighborhood-based differences in physical activity: an environment scale evaluation. Am J Public Health 93:1552–1558

Sallis JF (2000) Age-related decline in physical activity: a synthesis of human and animal studies. Med Sci Sports Exerc 32:1598–1600

Satterfield S, Cutler JA, Langford HG, Applegate WB, Borhani NO, Brittain E, Cohen JD, Kuller LH, Lasser NL, Oberman A, Rosner B, Taylor J, Vogt T, Walker G, Whelton PK (1991) Trials of hypertension prevention. Phase I design. Ann Epidemiol 1:455–471

Savastano S, Belfiore A, Di Somma C, Mauriello C, Rossi A, Pizza G, De Rosa A, Prestieri G, Angrisani L, Colao A (2010) Validity of bioelectrical impedance analysis to estimate body composition changes after bariatric surgery in premenopausal morbidly women. Obes Surg 20:332–339

Savva SC, Tornaritis M, Chadjigeorgiou C, Kourides YA, Savva ME, Panagi A, Chrictodoulou E, Kafatos A (2005) Prevalence and socio-demographic associations of undernutrition and obesity among preschool children in Cyprus. Eur J Clin Nutr 59:1259–1265

Schmidhuber J, Traill WB (2006) The changing structure of diets in the European Union in relation to healthy eating guidelines. Public Health Nutr 9:584–595

Schulze MB, Manson JE, Ludwig DS, Colditz GA, Stampfer MJ, Willett WC, Hu FB (2004) Sugar-sweetened beverages, weight gain, and incidence of type 2 diabetes in young and middle-aged women. JAMA 292:927–934

Seach KA, Dharmage SC, Lowe AJ, Dixon JB (2010) Delayed introduction of solid feeding reduces child overweight and obesity at 10 years. Int J Obes (Lond) 34:1475–1479

Segal L, Carter R, Zimmet P (1994) The cost of obesity: the Australian perspective. Pharmacoeconomics 5:45–52

Seidell JC, Cigolini M, Deslypere JP, Charzewska J, Ellsinger BM, Cruz A (1991) Body fat distribution in relation to physical activity and smoking habits in 38-year-old European men. The European Fat Distribution Study. Am J Epidemiol 133:257–265

Serdula MK, Ivery D, Coates RJ, Freedman DS, Williamson DF, Byers T (1993) Do obese children become obese adults? A review of the literature. Prev Med 22:167–177

Shields M, Tremblay MS (2010) Canadian childhood obesity estimates based on WHO, IOTF and CDC cut-points. Int J Pediatr Obes 5:265–273

Simmons G, Jackson R, Swinburn B, Yee RL (1996) The increasing prevalence of obesity in New Zealand: is it related to recent trends in smoking and physical activity? N Z Med J 109:90–92

Singh GK, Siahpush M, Kogan MD (2010) Neighborhood socio-economic conditions, built environments, and childhood obesity. Health Aff (Millwood) 29:503–512

Sjoberg A, Lissner L, Albertsson-Wikland K, Marild S (2008) Recent anthropometric trends among Swedish school children: evidence for decreasing prevalence of overweight in girls. Acta Paediatr 97:118–123

Soltoft F, Hammer M, Kragh N (2009) The association of body mass index and health-related quality of life in the general population: data from the 2003 Health Survey of England. Qual Life Res 18:1293–1299

Sorensen HT, Sabroe S, Rothman KJ, Gillman M, Fischer P, Sorensen TI (1997) Relation between weight and length at birth and body mass index in young adulthood: cohort study. BMJ 315:1137

Spiegel K, Leproult R, Van Cauter E (1999) Impact of sleep debt on metabolic and endocrine function. Lancet 354:1435–1439

Spiegel K, Tasali E, Penev P, Van Cauter E (2004) Brief communication: sleep curtailment in healthy young men is associated with decreased leptin levels, elevated ghrelin levels, and increased hunger and appetite. Ann Intern Med 141:846–850

Stamatakis E, Ekelund U, Wareham NJ (2007) Temporal trends in physical activity in England: the Health Survey for England 1991 to 2004. Prev Med 45:416–423

Stamatakis E, Zaninotto P, Falaschetti E, Mindell J, Head J (2010) Time trends in childhood and adolescent obesity in England from 1995 to 2007 and projections of prevalence to 2015. J Epidemiol Community Health 64:167–174

Stamler J, Wentworth D, Neaton JD (1986) Is relationship between serum cholesterol and risk of premature death from coronary heart disease continuous and graded? Findings in 356,222 primary screenees of the Multiple Risk Factor Intervention Trial (MRFIT). JAMA 256: 2823–2828

Staub K, Ruhli FJ, Woitek U, Pfister C (2010) BMI distribution/social stratification in Swiss conscripts from 1875 to present. Eur J Clin Nutr 64:335–340

Stettler N, Signer TM, Suter PM (2004) Electronic games and environmental factors associated with childhood obesity in Switzerland. Obes Res 12:896–903

Strauss RS, Knight J (1999) Influence of the home environment on the development of obesity in children. Pediatrics 103:e85

Strauss RS, Pollack HA (2003) Social marginalization of overweight children. Arch Pediatr Adolesc Med 157:746–752

Sun SS, Chumlea WC, Heymsfield SB, Lukaski HC, Schoeller D, Friedl K, Kuczmarski RJ, Flegal KM, Johnson CL, Hubbard VS (2003) Development of bioelectrical impedance analysis prediction equations for body composition with the use of a multicomponent model for use in epidemiologic surveys. Am J Clin Nutr 77:331–340

Sun G, French CR, Martin GR, Younghusband B, Green RC, Xie YG, Mathews M, Barron JR, Fitzpatrick DG, Gulliver W, Zhang H (2005) Comparison of multifrequency bioelectrical impedance analysis with dual-energy X-ray absorptiometry for assessment of percentage body fat in a large, healthy population. Am J Clin Nutr 81:74–78

Tang KH, Nguyen HH, Dibley MJ, Sibbritt DW, Phan NT, Tran TM (2010) Factors associated with adolescent overweight/obesity in Ho Chi Minh city. Int J Pediatr Obes 5:396–403

Taveras EM, Gillman MW, Kleinman K, Rich-Edwards JW, Rifas-Shiman SL (2010) Racial/ethnic differences in early-life risk factors for childhood obesity. Pediatrics 125:686–695

The NS, Adair LS, Gordon-Larsen P (2010) A study of the birth weight-obesity relation using a longitudinal cohort and sibling and twin pairs. Am J Epidemiol 172:549–557

The Trials of Hypertension Prevention Collaborative Research Group (1992) The effects of nonpharmacologic interventions on blood pressure of persons with high normal levels. Results of the Trials of Hypertension Prevention, Phase I. JAMA 267:1213–1220

The Trials of Hypertension Prevention Collaborative Research Group (1997) Effects of weight loss and sodium reduction intervention on blood pressure and hypertension incidence in overweight people with high-normal blood pressure. The Trials of Hypertension Prevention, Phase II. Arch Intern Med 157:657–667

Thomson M, Spence JC, Raine K, Laing L (2008) The association of television viewing with snacking behavior and body weight of young adults. Am J Health Promot 22:329–335

Tin Tin S, Woodward A, Thornley S, Ameratunga S (2009) Cycling and walking to work in New Zealand, 1991–2006: regional and individual differences, and pointers to effective interventions. Int J Behav Nutr Phys Act 6:64

Tjepkema M (2006) Adult obesity. Health Rep 17:9–25

Tochikubo O, Ikeda A, Miyajima E, Ishii M (1996) Effects of insufficient sleep on blood pressure monitored by a new multibiomedical recorder. Hypertension 27:1318–1324

Troiano RP, Briefel RR, Carroll MD, Bialostosky K (2000) Energy and fat intakes of children and adolescents in the United States: data from the national health and nutrition examination surveys. Am J Clin Nutr 72:1343S–1353S

Troiano RP, Berrigan D, Dodd KW, Masse LC, Tilert T, McDowell M (2008) Physical activity in the United States measured by accelerometer. Med Sci Sports Exerc 40:181–188

Tuck SP, Pearce MS, Rawlings DJ, Birrell FN, Parker L, Francis RM (2005) Differences in bone mineral density and geometry in men and women: the Newcastle Thousand Families Study at 50 years old. Br J Radiol 78:493–498

Tucker LA, Bagwell M (1991) Television viewing and obesity in adult females. Am J Public Health 81:908–911

Tucker LA, Friedman GM (1989) Television viewing and obesity in adult males. Am J Public Health 79:516–518

Tunceli K, Li K, Williams LK (2006) Long-term effects of obesity on employment and work limitations among U.S. Adults, 1986 to 1999. Obesity (Silver Spring) 14:1637–1646

Ulijaszek SJ, Kerr DA (1999) Anthropometric measurement error and the assessment of nutritional status. Br J Nutr 82:165–177

United States Department of Health and Human Services (1996) Physical activity and health: a report of the Surgeon General Executive Summary. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention; President's Council on Physical Fitness and Sports, Atlanta

United States Department of Health and Human Services (2000) Healthy people 2010. U.S. Department of Health and Human Services, Washington, DC

United States Department of Health and Human Services (2008) 2008 physical activity guidelines for Americans: be active, healthy, and happy! U.S. Department of Health and Human Services, Washington, DC

van der Kooy K, Seidell JC (1993) Techniques for the measurement of visceral fat: a practical guide. Int J Obes Relat Metab Disord 17:187–196

van Dijk L, Otters HB, Schuit AJ (2006) Moderately overweight and obese patients in general practice: a population based survey. BMC Fam Pract 7:43

Van Dyck D, Cardon G, Deforche B, Sallis JF, Owen N, De Bourdeaudhuij I (2010) Neighborhood SES and walkability are related to physical activity behavior in Belgian adults. Prev Med 50(Suppl 1):S74–S79

Viner RM, Cole TJ (2005) Television viewing in early childhood predicts adult body mass index. J Pediatr 147:429–435

Wagner DR, Heyward VH (2000) Measures of body composition in blacks and whites: a comparative review. Am J Clin Nutr 71:1392–1402

Wang J, Thornton JC, Bari S, Williamson B, Gallagher D, Heymsfield SB, Horlick M, Kotler D, Laferrere B, Mayer L, Pi-Sunyer FX, Pierson RN Jr (2003) Comparisons of waist circumferences measured at 4 sites. Am J Clin Nutr 77:379–384

Wang YC, Bleich SN, Gortmaker SL (2008) Increasing caloric contribution from sugar-sweetened beverages and 100% fruit juices among US children and adolescents, 1988–2004. Pediatrics 121:e1604–e1614

Wardle J, Waller J, Jarvis MJ (2002) Sex differences in the association of socio-economic status with obesity. Am J Public Health 92:1299–1304

Weyermann M, Brenner H, Rothenbacher D (2007) Adipokines in human milk and risk of overweight in early childhood: a prospective cohort study. Epidemiology 18:722–729

Whitworth JA (2003) 2003 World Health Organization (WHO)/International Society of Hypertension (ISH) statement on management of hypertension. J Hypertens 21:1983–1992

WHO Multicentre Growth Reference Study Group (2006) Enrolment and baseline characteristics in the WHO Multicentre Growth Reference Study. Acta Paediatr Suppl 450:7–15

Willett W (1988) Nutritional epidemiology. Oxford University Press, New York

Willett K, Jiang R, Lenart E, Spiegelman D, Willett W (2006) Comparison of bioelectrical impedance and BMI in predicting obesity-related medical conditions. Obesity (Silver Spring) 14:480–490

Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB (1998) Prediction of coronary heart disease using risk factor categories. Circulation 97:1837–1847

Wolf AM, Colditz GA (1998) Current estimates of the economic cost of obesity in the United States. Obes Res 6:97–106

World Health Organization (2000) Obesity: preventing and managing the global epidemic. Report of a WHO consultation. World Health Organ Tech Rep Ser 894:i–xii, 1–253

World Health Organization (2002) World Health Organization's World Health Report, Reducing risks, promoting healthy life. http://www.who.int/whr/2002/en/whr02_en.pdf. Accessed 15 Jan 2011

World Health Organization (2010) Global database on body mass index: an interactive surveillance tool for monitoring nutrition transition 2010. http://apps.who.int/bmi/index.jsp. Accessed 10 June 2010

Wright JD, Kennedy-Stephenson J, Wang CY, McDowell MA (2004) Trends in intake of energy and macronutrients-United States, 1971–2000. MMWR Morb Mortal Wkly Rep 53:80–82

Wright CS, Rifas-Shiman SL, Rich-Edwards JW, Taveras EM, Gillman MW, Oken E (2009) Intrauterine exposure to gestational diabetes, child adiposity, and blood pressure. Am J Hypertens 22:215–220

Wrotniak BH, Shults J, Butts S, Stettler N (2008) Gestational weight gain and risk of overweight in the offspring at age 7 y in a multicenter, multiethnic cohort study. Am J Clin Nutr 87: 1818–1824

Zimmerman FJ, Bell JF (2010) Associations of television content type and obesity in children. Am J Public Health 100:334–340

Epidemiology of Respiratory Allergies and Asthma

56

Jeroen Douwes and Neil Pearce

Contents

J. Douwes (✉)
Centre for Public Health Research, School of Public Health, Massey University Wellington, Wellington, New Zealand

N. Pearce
Department of Medical Statistics, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

Centre for Public Health Research, School of Public Health, Massey University Wellington, Wellington, New Zealand

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_50,

56.1 Introduction

Asthma has puzzled and confused physicians from the time of Hippocrates to the present day. The word “asthma” comes from a Greek word meaning “panting” (Keeney 1964), but reference to asthma can also be found in ancient Egyptian, Hebrew, and Indian medical writings (Ellul-Micallef 1976; Unger and Harris 1974). There were clear observations of patients experiencing attacks of asthma in the second century and evidence of disordered anatomy in the lung as far back as the seventeenth century (Dring et al. 1689).

More recently, Western countries have experienced an epidemic of respiratory allergy and asthma prevalence (and incidence) that appears to have commenced after the Second World War and has only recently peaked and begun to decline (Pearce et al. 2007). The reasons for the decline remain as mysterious as the reasons for the epidemic itself (Weiland and Pearce 2004). Current WHO estimates suggest that some 300 million persons suffer from asthma, which is the most common chronic disease among children (WHO 2007). In 2010, the annual economic cost of asthma in the United States was estimated to be >$15 billion in direct costs and >$5 billion in indirect costs such as lost productivity (American Lung Association 2010).

Formal epidemiological studies of respiratory allergies and asthma commenced about 50 years ago, and in some respects they have failed to deliver the insight or testable hypotheses that have derived from studies of other common diseases. The roots of this failure lie in the problem of defining asthma, and the consequent difficulty of reproducible and conclusive case ascertainment; in the transient nature of the principal symptoms, the fact that asthma is heterogeneous in its underlying pathophysiology; and in the absence of simple sensitive and specific markers for the condition. However, the epidemiology of asthma and allergy is currently undergoing a rapid expansion and has shifted attention from traditional risk factors which may exacerbate asthma, to protective factors that may prevent respiratory allergies and asthma. This growth of interest is occurring in the context of major concerns about the increasing burden of respiratory allergy and asthma morbidity, and the realization that previous etiological theories based on animal models of asthma and clinical observations do not appear to explain the large global increases in asthma prevalence.

This chapter begins with the definitions and the underlying pathology and physiology of respiratory allergies and asthma, before considering the distinctive features of asthma epidemiology. We subsequently review the global burden of respiratory allergies and asthma and discuss a wide range of potential risk and protective factors. The chapter concludes with a discussion of the role of epidemiology in respiratory allergies and asthma research, and the challenges it is facing, followed by a discussion of the foremost issues for the future of respiratory allergies and asthma epidemiology.

56.2 Definitions of Respiratory Allergies and Asthma

56.2.1 Respiratory Allergy

Allergy can be defined as adverse acute or chronic hypersensitivity reactions resulting from immunologic sensitization with production of immunoglobulin (Ig) E against a specific agent or allergen. Thus, the term "allergy" refers to symptomatic conditions (allergic asthma, rhinitis, etc.), whereas the term "sensitization" refers to an individual's immune status assessed by in vivo or in vitro diagnostic tests (see Sect. 56.4.3). Symptoms can be induced by inhalation of allergens, even at very low concentrations. Individuals that are not sensitized to these allergens will usually not show symptoms even with very high exposure. Symptoms in sensitized subjects are caused by inflammatory reactions initiated by allergen-specific IgE antibodies present in the airways. Only a proportion of sensitized subjects show symptoms and are thus also allergic. It can take weeks to years between the first encounter with an allergen and the development of an allergy.

56.2.2 Atopy

"Atopy" (allergic sensitization) is a common term for IgE-mediated sensitization and/or allergic reactions. In population studies, the term "atopy" is used to indicate the predisposition of individuals to produce increased levels of specific or total IgE after exposure to common allergens such as house dust mite, pet, and various food allergens. It is usually assessed by skin prick tests or specific and/or total serum IgE against common allergens (see Sect. 56.4.3), and it can therefore be defined either in terms of skin prick text positivity or elevated serum IgE levels (Pearce et al. 1999). Depending on the definition, about 20% to 40% of people in affluent countries are atopic. In population studies, atopy is often associated with an increased risk of asthma (Pearce et al. 1999), but the association is stronger in "Westernized" countries than in developing countries (Weinmayr et al. 2007).

56.2.3 Aeroallergens

Many macromolecules (particularly proteins) of non-human origin, including those of animals (e.g., arthropod proteins, animal danders, proteins in excretia), plants (e.g., pollens, latex dust), and microorganisms (e.g., spores of fungi such as *Alternaria, Aspergillus,* and *Penicillium*), can act as allergens by inducing a specific IgE response and provoke allergic reactions in sensitized subjects.

Dust mites produce the predominant inhalant allergens in many parts of the world. The most common mite species that produce allergens are *Dermatophagoides pteronyssinus* and *D. farinae*. The major allergens produced by *D. pteronyssinus* (called Der p 1 and Der p 2) are proteases present in high amounts in fecal pellets. *D. farinae* produces as its major allergen Der f 1. Elevated levels of these allergens have been detected in house dust, mattress dust, and bedding in damp homes (van Strien et al. 1994).

Other important inhalant allergens include proteins associated with cats and dogs, cockroaches, grass and tree pollens, and fungi such as *Alternaria*. Allergens in the occupational environment can range from cow urinary proteins in farming situations to fungal enzymes in the biotechnology and bakery industry. Low molecular weight chemicals such as diisocyanates (e.g., tolueen diisocyanate, TDI) can also cause occupational allergic asthma, but the specific immunological mechanisms have not yet been resolved.

56.2.4 Asthma

The definition of asthma initially proposed at the Ciba Foundation conference in 1959 (Ciba Foundation Guest Symposium 1959) and endorsed by the American Thoracic Society in 1962 (American Thoracic Society Committee on Diagnostic Standards 1962) is that "asthma is a disease characterized by wide variation over short periods of time in resistance to flow in the airways of the lung." Although these features receive lesser prominence in some current definitions, as the importance of airways inflammation is appropriately recognized, they still form the basis of the recent Global Initiative for Asthma (GINA) description of asthma as:

> . . . a chronic inflammatory disorder of the airways in which many cells and cellular elements play a role. The chronic inflammation is associated with airway hyper-responsiveness that leads to recurrent episodes of wheezing, breathlessness, chest tightness, and coughing, particularly at night or in the early morning. These episodes are usually associated with widespread, but variable, airflow obstruction within the lung that is often reversible either spontaneously or with treatment. (GINA 2006)

These three components, chronic airways inflammation, reversible airflow obstruction, and enhanced bronchial reactivity, form the basis of current definitions of asthma. They also represent the major pathophysiological events leading to the symptoms of wheezing, breathlessness, chest tightness, cough, and sputum by which physicians clinically diagnose this disorder.

56.2.5 Clinical Asthma

There is no single test or pathognomic feature which defines the presence or absence of asthma. Furthermore, the variability of the condition means that evidence of it may or may not be present on the day, or at the time, that someone is assessed. Thus, a diagnosis of asthma is made on the basis of the clinical history, combined with physical examination and respiratory function tests over a period of time. Several studies have found the prevalence of physician-diagnosed asthma to be substantially lower than the prevalence of asthma symptoms in the community (e.g., Asher et al. 1998). This is not surprising since a clinical diagnosis of asthma can only be made if a person presents him or herself to a doctor. This requires an initial self-assessment of the symptoms (in terms of severity and frequency), as well as access to a doctor

Table 56.1 GINA classification of asthma severity

Asthma classification	Criteria
Intermittent	Symptoms less than once a week; brief exacerbations; nocturnal symptoms not more than twice a month; FEV_1 or PEF $\geq$80% predicted; FEV_1 or PEF variability <20%
Mild persistent	Symptoms more than once a week but less than once a day; exacerbations may affect activity and sleep; nocturnal symptoms more than twice a month; FEV_1 or PEF $\geq$80% predicted; FEV_1 or PEF variability 20–30%
Moderate persistent	Symptoms daily; exacerbations may affect activity and sleep; nocturnal symptoms more than once a week; daily use of inhaled short-acting β_2-agonist; FEV_1 or PEF 60–80% predicted; FEV_1 or PEF variability >30%
Severe persistent	Symptoms daily; frequent exacerbations; frequent nocturnal asthma symptoms; limitation of physical activities; FEV_1 or PEF $\leq$60% predicted; FEV_1 or PEF variability >30%

FEV_1 Forced Expiratory Volume in one second, *PEF* Peak Expiratory Flow

once a self-assessment has been made. Several medical consultations may then be required. Thus, diagnosed asthma is dependent not only on morbidity but also on patient perceptions, physician practice, and the availability of health care.

There are a number of tests that may facilitate the diagnosis and monitoring of asthma. Measurements of lung function are the most frequently used and provide important information on airflow limitation including variability, reversibility, and severity. Airflow limitation is most often measured using spirometry or a peak (expiratory) flow (PEF) meter. Spirometry is the preferred method and forced expiratory volume in one second (FEV_1), and peak expiratory flow (PEF) can be derived from this method as indicators for airway obstruction. Pre- and postbronchodilator treatment is important since it will establish whether obstruction is reversible and will distinguish it from chronic obstructive pulmonary disease (COPD) in which obstruction is mostly irreversible. Reversibility of airway obstruction defined as FEV_1 of $\geq$12% and $\geq$200mL from the prebronchodilator value is generally accepted as a valid indication of asthma (GINA 2006). However, due to the highly variable nature of the condition, repeated lung function tests are required. PEF meters are inexpensive and easy to use, but they are less precise than spirometry and may underestimate the degree of airflow limitation (Aggarwal et al. 2006).

In subjects with asthma symptoms but normal lung function, bronchial hyperresponsiveness (BHR) testing is often used as a diagnostic aid. BHR constitutes airway narrowing to non-specific stimuli, such as exercise, cold air, and chemical irritants, and can be measured as airway responsiveness to histamine, methacholine, adenosine-5′-monophosphate (AMP), hypertonic saline, or exercise challenge (de Meer et al. 2004a). However, although BHR is related to asthma, it may occur independently of asthma, and vice versa (Pearce et al. 1998a), which makes this test of limited use for individual asthma-diagnostics (see Sect. 56.4.1).

The degree of asthma severity is commonly classified using GINA criteria (GINA 2006) which subdivide asthma into four categories based on frequency and severity of symptoms and lung function (with the worst feature determining the severity classification) (see Table 56.1). However, as emphasized in the updated

2010 GINA guidelines (GINA 2010), asthma involves both the severity of the disease and its responsiveness to treatment. Therefore, asthma could present with severe symptoms and poor lung function but could become completely controlled with low-dose treatment. Thus, the classification of severity had poor predictive value for what treatment was required and what the response to treatment might be. As a consequence, the 2010 guidelines do no longer include the 2006 asthma severity classification. However, it is still a commonly used classification in many epidemiological studies.

56.2.6 Defining Asthma in Epidemiological Surveys

Defining asthma in population-based epidemiological surveys of asthma prevalence or incidence poses even greater difficulties than defining asthma in individual patients. As a result, comparisons of diagnosed asthma between populations are fraught with difficulty, as the differences in diagnostic practice may be greater in magnitude than the real differences in asthma morbidity.

Thus, asthma prevalence surveys usually focus on self-reported (or parental reported) "asthma symptoms" rather than diagnosed asthma. Standardized questionnaires on asthma symptoms have therefore become the cornerstone of large studies of the incidence or prevalence of asthma (Burney et al. 1994; Asher et al. 1995). This approach allows a large number of participants to be surveyed without great cost, in a short time period. Wheezing, chest tightness, breathlessness, and coughing are all symptoms clinically associated with asthma, but epidemiological studies have shown that wheezing is the most important symptom for the identification of asthma, and the majority of questionnaires used to assess asthma prevalence are based on this symptom (Pearce et al. 1998a) (see Sect. 56.4.1 for more detail).

An alternative approach to symptom questionnaires has been to use more "objective" measures such as bronchial responsiveness testing, either alone or in combination with questionnaires (Toelle et al. 1992). However, some have questioned the validity of BHR for the assessment of asthma (Pearce et al. 1998a) (see Sect. 56.4 for more detail).

56.3 Mechanisms of Respiratory Allergies and Asthma

56.3.1 Asthma Phenotypes

Ten years ago it was widely believed that asthma was an atopic disease caused by allergen exposure. The fundamental etiological mechanism was that allergen exposure, particularly in infancy, produced atopic sensitization, and continued exposure resulted in asthma through the development of eosinophilic airways inflammation, bronchial hyperresponsiveness and reversible airflow obstruction. In recent years, it has become increasingly evident that this picture is, at best, too simplistic (Ronchetti et al. 2007; Weinmayr et al. 2007). A systematic review of population-based studies (Pearce et al. 1999) has shown that the proportion of asthma cases that are attributable to atopy (defined as skin prick test positivity) is usually less than one-half. Standardized comparisons across populations or time periods also show only weak and inconsistent associations between the prevalence of asthma and the prevalence of atopy. For instance, a comparison of asthma and atopy in 9- to 11-year-olds in Albania and the United Kingdom (Priftanji et al. 2001) showed large differences in the prevalence of current wheeze (4.4% and 9.7%, respectively) and exercise-induced bronchial reactivity (0.8% and 5.4%), but not in skin prick test positivity (15.0% and 17.8%), suggesting that large variations in asthma prevalence can occur without differences in frequency of atopy. This was confirmed by the International Study of Allergies and Asthma in Children (ISAAC; see Sects. 56.5.3 and 56.5.4) which showed that the association between atopy and asthma symptoms differed strongly among populations but increased with the level of economic development (Weinmayr et al. 2007). In this study, the proportion of current wheeze attributable to atopy ranged from 0% in Ankara (Turkey) to 93.8% in Guangzhou (China); the overall proportion of asthma cases that were attributable to atopy was only 41% in affluent countries and 20% in non-affluent countries. This is consistent with a recent study among 3,960 children in Ecuador which found a population attributable fraction of only 2.4% for recent wheeze and atopy (Moncayo et al. 2010). A similar study in Salvador, Brazil suggested the proportion of asthma attributable to atopy to be 24.5% (Souza da Cunha et al. 2010), which, although higher than in Ecuador, is substantially lower than the 50% in most Western countries. The European Community Respiratory Health Survey (see Sects. 56.5.2 and 56.5.4) found that the proportion of asthma attributable to atopy in adults ranged from 4% to 61% between individual study centers with an overall estimate of only 30% for all centers combined (Sunyer et al. 2004). Finally, time trend studies have shown different patterns for atopy and astma in the same geographical areas. For example, a recent study in Sweden in 7- to 8-year-olds showed a significant increase in allergic sensitization over a 10-year period (1996–2006) despite asthma symptoms remaining stable over the same period (Bjerg et al. 2010).

Recent studies using sputum induction and/or bronchoalveolar lavage (BAL) techniques to measure and characterize airways inflammation in asthmatics have also demonstrated that less than 50% of asthma cases are attributable to eosinophilic airway inflammation, the hallmark of allergic asthma (Douwes et al. 2002a; Simpson et al. 2006). Thus, evidence from studies of eosinophilia and asthma is consistent with that from studies of atopy and asthma: in both instances, at most about one-half of asthma cases appears to be due to "allergic" mechanisms. This further adds to the evidence that allergic mechanisms may not be the only, or the most important, underlying mechanism for asthma.

56.3.2 Immunology

56.3.2.1 Allergic Asthma

Allergic asthma is caused by IgE-mediated inflammatory mechanisms in which a large number of cells play a role including mast cells, eosinophils, T lymphocytes, dendritic cells, and macrophages. Briefly, the sensitization process involves the adaptive (or acquired) immune system whereby allergens interact with dendritic cells in the airway mucosa which migrate to the regional lymph nodes where the allergens are presented to B and T cells. This results (through T-helper-2 (Th_2) responses) in the production of allergen-specific IgE. Once allergic, a subject can develop symptoms minutes after being exposed. This is known as the early phase allergic reaction, and symptoms develop as a result of mast cell degranulation and release of inflammatory mediators through allergen IgE antibody complexes at the surface of the mast cells, causing contraction of bronchial smooth muscle and edema in the airways. Clinically, this results in a decreased lung function and symptoms of wheeze, shortness of breath, chest tightness, and coughing.

During the late phase of the allergic reaction (4 to 8 h after exposure), eosinophil-related inflammatory reactions are particularly important. A critical step in this late phase reaction is the activation of Th_2 cells which release several proinflammatory cytokines including IL-5 resulting in the influx and activation of eosinophils. This reaction is characterized by the development of a non-specific BHR that can continue for several days. Repeated exposures can result in more permanent BHR.

56.3.2.2 Non-Allergic Asthma

As noted above, a substantial proportion of all asthma cases have an underlying pathology that is different from that observed in “classic” allergic asthma (Douwes et al. 2002a; Simpson et al. 2006). Patients may have severe and persistent asthma in the absence of eosinophilic inflammation and may experience an exacerbation of asthma without an increase in eosinophilic inflammation (Turner et al. 1995). Repeated assessments of airway inflammation over time have shown that the non-eosinophilic asthma phenotype is reproducible both in the short (4 weeks) and long term (1 to 5 years) (Simpson et al. 2006). However, the underlying mechanisms of non-eosinophilic/non-allergic asthma are not fully understood. One study in 93 non-smoking adult asthmatics found elevated sputum airway eosinophilia in 41% of all asthma cases, 20% had elevated levels of neutrophils, 8% had a mixed inflammatory profile with both cell types being elevated, and the remainder (31%) had no signs of airway inflammation with eosinophil and neutrophil levels both being within the normal range (Simpson et al. 2006). This suggests that asthma can be categorized into four inflammatory subtypes based on the sputum eosinophil and neutrophil proportions: eosinophilic asthma, neutrophilic asthma, mixed granulocytic asthma, and paucigranulocytic asthma (Simpson et al. 2006).

The common pathophysiological features of neutrophilic asthma involve an IL-8-mediated neutrophil influx, and the subsequent neutrophil activation is a potent stimulus to increased airway hyperresponsiveness (Simpson et al. 2007). Although the stimuli that trigger this response are diverse (endotoxin, ozone, particulates, virus infection), the common features are consistent with activation of innate immune mechanisms (involving Toll-like receptors and CD14) rather than IgE-mediated activation of acquired immunity.

There is also the potential for combined activation of both innate and allergen-specific inflammatory mechanisms in asthma. This may be the case in mixed granulocytic asthma and may explain the ability of ozone and NO2 to potentiate allergen-induced asthmatic responses (Jenkins et al. 1999). The pathophysiological mechanisms involved in paucigranulocytic asthma are not clear.

Clinically, the eosinophilic and non-eosinophilic phenotypes appear very similar with only small differences in lung function, airway hyperreactivity, corticosteroid use, and β_2 agonist-induced reversibility in FEV_1 (Simpson et al. 2006; Berry et al. 2007). However, there are also distinct differences: non-eosinophilic asthmatics appear to be less atopic, have normal subepithelial layer thickness, and perhaps most importantly, they have a poor short-term response to treatment with inhaled corticosteroids (Berry et al. 2007). Thus, despite clinical similarities they represent distinct pathological phenotypes.

Non-allergic occupational asthma is also very common. For instance, in many occupational environments where workers are exposed to organic dust (e.g., farmers, grain workers), the majority of asthma cases are non-IgE-mediated and are related to chronic exposure to environmental irritants. The underlying inflammatory mechanisms involve innate immune responses which are often directed against constituents of bacteria and fungi (Douwes et al. 2002a, b).

56.4 How to Measure Respiratory Allergies and Asthma in Epidemiological Studies

56.4.1 Measuring Incidence and Prevalence

Ideally, we would wish to measure incidence in epidemiological studies of asthma. However, in practice, asthma incidence is very difficult to measure, both because of the intensive long-term monitoring required and because of the difficulty of establishing the date of onset of the condition. Therefore, most studies involve prevalence rather than incidence. Similar techniques to those described in this section can be used in incidence studies, although incidence studies may involve more intensive monitoring to enable a diagnosis of asthma to be more firmly established. Asthma prevalence reflects both the incidence of asthma and the average duration of the condition. Thus, a population may have a high prevalence of asthma either due to a high exposure to factors (genetic or environmental) which induce asthma or because of high exposure to factors which incite, exacerbate, or prolong asthma symptoms in those who have previously developed the disease (Dolovich and Hargreave 1981).

56.4.1.1 Diagnosed Asthma

Although asthma can be conceptualized either in terms of symptoms such as wheezing or in terms of the underlying bronchial inflammation, the essential feature of asthma (at least in clinical and epidemiological terms) is variable airflow obstruction which can be reversed by treatment or is self-limiting (see Sect. 56.2.4). This poses several problems with the use of “diagnosed asthma” in asthma prevalence studies, since the diagnosis of “variable airflow obstruction” usually requires several medical consultations over an extended period. It is therefore not surprising that several studies have found the prevalence of physician-diagnosed asthma to be substantially lower than the prevalence of asthma symptoms. For example, Ehrlich et al. (1995), in a survey in school children in Cape Town, found that among children with more than 12 attacks of wheezing in the previous 12 months, only 60% were reported as asthmatic and only 55% as receiving regular treatment.

These problems with using "diagnosed asthma" as an outcome measure in epidemiological studies not only affect prevalence estimates but also affect time trends (Hill et al. 1989) and geographical and social patterns of asthma prevalence. For example, some studies (e.g., Peckham and Butler 1978) have found diagnosed asthma to be more common in upper social class children, possibly because of greater access to health care and therefore greater likelihood of being labeled as asthmatic. The same concerns apply, perhaps to a greater extent, in surveys of use of asthma medication or health services. Although such information may be of value in asthma morbidity studies, it is of very limited use in prevalence studies (Pearce and Beasley 1999).

These issues are of particular concern in geographical prevalence comparisons since there are major international and regional differences in access to health care and labeling of asthma. Thus, some international comparisons have found differences in diagnosed asthma to be much greater than differences in reported asthma symptoms (e.g., Pearce et al. 1993), these differences in diagnosed asthma may partially reflect differences in access to health services and diagnostic practice rather than genuine differences in asthma prevalence. For example, Dodge and Burrows (1980) suggest that "the epidemiology of asthma is a reflection of the diagnostic habits of physicians in the locale, as well as an indicator of the frequency of a specific syndrome."

Such problems of differences in diagnostic practice could be minimized by using a standardized protocol for asthma diagnosis in prevalence studies. However, this is rarely a realistic option since it requires repeated contacts between the study participants and physicians, and this is often not possible or affordable in large-scale epidemiological studies. Although self-reported histories of physician-diagnosed asthma have been found to be relatively valid, this relates to diagnosed asthma rather than true asthma prevalence.

56.4.1.2 Symptoms

Questionnaires on asthma symptoms are the cornerstone of large-scale epidemiological surveys of asthma prevalence. These have the advantage of being inexpensive and simple to administer to large numbers of participants on a single day and will discriminate those with variable airflow obstruction that causes noticeable symptoms. The key issue is that questionnaires should obtain information in a standardized manner on symptoms which are directly related to "variable airflow obstruction." This definition of asthma implies a condition in which symptoms occur from time to time, rather than the presence or absence of symptoms on a particular day. Thus, operational definitions of asthma involve specification of the time period during which symptoms may have occurred. In particular, “current symptoms” are usually defined as symptoms at any time in the previous 12 months. Although there has been concern that repeated questioning of subjects could increase awareness of respiratory symptoms, a comparison of findings from an intensive longitudinal study and a prevalence study within similar populations found very similar symptom prevalence suggesting that repeated questioning of the longitudinal study population had not biased the prevalence estimates (Sears et al. 1997).

Standard written questionnaires have been the principal instrument for measuring asthma symptom prevalence in community surveys, and in homogeneous populations these have been standardized, validated, and shown to be reproducible (Burney et al. 1989). A number of symptoms including wheezing, chest tightness, breathlessness, and coughing with or without sputum are recognized by physicians as indicative of asthma. Of these the most important symptom for the identification of asthma in epidemiological studies is wheezing, and most questionnaires have focused on this.

Symptoms may be absent despite variable airflow obstruction (asthma), and symptoms may also occur in the absence of asthma. Thus, wheezing is not synonymous with diagnosed asthma; in some instances, wheezing may indicate asthma which has not been diagnosed, but it may also indicate other diseases (particularly chronic bronchitis and emphysema in persons aged 45 years or more) or may be unrelated to other symptoms or disease; conversely, diagnosed asthmatics may not experience or recognize wheeze as a symptom. Nevertheless, wheezing remains the symptom which is most characteristic of asthma, particularly in persons aged 5 to 34 years where asthma is less likely to be confused with bronchitis or emphysema. Thus, most asthma symptom prevalence questionnaires focus on “current wheezing” (defined as wheezing at any time in the previous 12 months), but they usually also include additional questions on the frequency of wheezing and the circumstances in which wheezing occurs (wheezing while at rest, wheezing after exercise, wheezing in the absence of a cold, waking with wheezing), as well as questions on related symptoms (e.g., waking with cough or severe episodes of breathlessness).

A large number of such questions have been used in epidemiological surveys in the last decades. In particular, the questionnaires for the European Community Respiratory Health Survey (ECRHS) in adults (Burney et al. 1994) and the International Study of Asthma and Allergies in Childhood (ISAAC) (Asher et al. 1995) have now become established as the standard questionnaires for use in prevalence surveys.

One of the major problems with the use of written questionnaires in international comparisons is with their translation into the different languages of the communities studied. This is particularly problematic for language groups (e.g., German, French, Dutch) which have no colloquial term for “wheezing.” To ensure that data regarding the frequency and severity of symptoms is comparable in international comparisons of asthma prevalence, it is important that the questionnaire is translated in a manner that ensures that the terms describing symptoms correspond as closely as possible to the international recommended questions.

56.4.1.3 Physiological Measures

The problems of validity, repeatability, and translation of questionnaires have led to attempts to find more "objective" physiological measures for use in prevalence studies. These measures carry with them their own set of limitations both with regard to their "objectivity" (i.e., they are not unequivocally measures of asthma prevalence) and their practicality in large population-based surveys. Nevertheless, they form a useful complement to symptom questionnaires.

Diminished lung function, besides being used as an outcome measure in prevalence studies, is also a risk factor for the development of wheezing and/or asthma (Martinez et al. 1988). Lung function measurements include forced expiratory volume in one second (FEV_1), forced vital capacity (FVC), and peak expiratory flow (PEF). These are all forced expiratory maneuvers undertaken following a maximum inspiratory maneuver. While maximum inspiratory and expiratory maneuvers may lead to small transient changes in airway calibre, the simplicity, standardization, and reproducibility of such maximal expiratory manoeuvres (particularly FEV_1 and FVC) make them the measurements of choice in epidemiological studies (Quanjer et al. 1993), and standardized guidelines are available (American Thoracic Society 1987).

However, such measures may have limited use in prevalence studies, since the airflow obstruction in asthma is reversible over short periods of time and may not be present on the day of assessment. Indeed, one of the characteristic features of reversible airflow obstruction in asthma is its diurnal variation, in that the degree of airflow obstruction is greatest at night or on wakening and least during the day when lung function testing is likely to be undertaken. Thus, a significant proportion of asthmatics may have lung function within normal limits on any one day, and not exhibit the 15% postbronchodilator improvement in FEV_1 or PEF values that is required for a physiological diagnosis of asthma to be made.

A better approach is to undertake repeat measurements of lung function before and after bronchodilator, at different times of the day over a period of time (e.g., two weeks); however, this is difficult and expensive in practice. It is this variation and reversibility of changes in lung function which typifies asthma, and this cannot be captured in a one-off prevalence survey on a particular day. Similarly, a one-off clinical examination with auscultation of the chest to identify the proportion of subjects with wheeze is also of limited value.

Measurements of bronchial hyperresponsiveness (BHR) in epidemiological studies of obstructive respiratory disease have been increasingly used in the past decades, and standard procedures are now well established (James and Ryan 1997). Bronchial hyperresponsiveness (BHR) testing was originally developed in a clinical setting and has conventionally been defined in terms of the dose of agonist producing a 15 or 20% fall in FEV_1. Although BHR is often defined in terms of a dichotomy (e.g., subjects are defined as having BHR if they experience a 15 or 20% fall in FEV_1 at a dose of agonist below a specified level), it can also be regarded as a continuous measure using the dose-response slope (de Meer et al. 2005).

BHR tests are neither wholly sensitive nor specific for asthma (Pekkanen and Pearce 1999), and recent results of the ISAAC Phase II study (see Sect. 56.5.3) involving BHR testing in 6,826 school children in 16 countries showed that, although BHR and asthma symptoms were associated within individual centers, a poor correlation was found between the level of BHR and wheeze across all participating centers ($\rho = 0.23$, $p = 0.294$) and high prevalences of BHR were not confined to centers with high prevalences of asthma symptoms (Buchele et al. 2010). The same study also showed that BHR (assessed by using hypertonic saline) was more closely associated with atopic asthma with a significant correlation between center-specific mean level of BHR and wheeze in atopic ($\rho = 0.64$, $p = 0.002$), but not in non-atopic children ($\rho = 0.21$, $p = 0.35$). Also, several studies have reported an absence of BHR in a substantial proportion of subjects with asthma, and BHR has also been reported in those without asthma (Josephs et al. 1990). BHR also occurs in chronic bronchitis and in tobacco smokers (Burney et al. 1987).

Perhaps the fundamental problem is that persons with non-specific BHR may have a greater tendency to experience asthma symptoms if they are exposed to the relevant stimulus in the laboratory but whether they do actually experience asthma symptoms in daily life, will depend on whether exposure occurs and at what level. Thus, the severity of airflow obstruction may be determined by the interaction of non-specific reactivity and the strength of a provoking bronchoconstrictor stimulus (Josephs et al. 1990). Furthermore, BHR is only one mechanism contributing to the clinical expression of variable airflow obstruction and may contribute to a greater or lesser extent in different individuals and within the same individual at different times (Josephs et al. 1990).

Thus, although BHR is related to asthma and may be involved in many of the pathways by which variable airflow obstruction may occur, variable airflow obstruction may occur independently of BHR, and vice versa (Pearce et al. 2000a). Also, using BHR to define asthma as has been suggested by some (Toelle et al. 1992), risks not only introducing error but will also bias the asthma sample toward one particular asthma phenotype (i.e., atopic asthma). Thus, asthma and BHR remain separate phenomena which both may involve inflammation of the airways, and which are both worthy of study in their own right.

56.4.2 Measuring Risk Factors

Asthma epidemiology studies typically involve measuring the effect of an exposure (e.g., air pollution) on a particular outcome (e.g., asthma, severe asthma, asthma hospital admission, asthma death) in a population. Strictly speaking, the term *exposure* refers to the presence of a substance (e.g., house dust mite allergen) in the external environment, whereas the term *dose* refers to the amount of substance that reaches susceptible targets within the body, such as the airways. In some situations (e.g., in a coal mine), measurements of external exposures may be strongly correlated with internal dose, whereas in other situations (e.g., indoor exposure to house dust mite allergen), the dose may depend on individual lifestyle and activities and may therefore be only weakly correlated with the environmental exposure levels. However, in this chapter we use the term “exposure” in the very general sense to denote any attribute or agent which may increase the risk of experiencing the outcome under study. It includes demographic factors (age, sex, ethnicity, social class) and genetic factors as well as environmental risk factors (measured externally and/or internally).

The environmental exposures most commonly associated with asthma (development or exacerbation) in the general population include a wide range of bioaerosols such as non-allergenic microbial agents, including bacterial endotoxin, house dust mite, pet, and cockroach allergens, and allergenic and non-allergenic fungal agents (see Sect. 56.6.1). Also, more than 250 agents have been identified as causes of occupational asthma (see Sect. 56.6.1) with most of these being biological in origin.

Exposure levels can be assessed with regard to the *concentration* of the substance in the environment (e.g., allergen concentration in the air) and the *duration* of time for which exposure occurs. The risk of developing sensitization to an allergen may be much greater if the duration of exposure is long and/or the exposure is high, and the total *cumulative exposure* may therefore be important. However, once a person has become sensitized, then the concentration may be crucial in provoking an acute attack, and such attacks may occur after exposure of a relatively short duration. For protracted etiological processes, the time pattern of exposure may be important, and it is possible to assess this by examining the separate effects of exposures in various time windows prior to the occurrence and recognition of clinical disease (Pearce 1992). For example, several studies have found that exposure to environmental tobacco smoke in the first years of life may be as relevant as current exposures with regard to current asthma symptoms (e.g., Shaw et al. 1994). Similarly, it has been suggested that occupational asthma is most likely to occur after about 1 to 3 years of exposure to a sensitizing agent (Anto et al. 1996).

Epidemiological studies rarely have optimal exposure/dose data and often rely on relatively crude measures of exposure. The key issue is that the exposure data need not be perfect but that it must be of similar quality for the various groups being compared. Provided that this principle is followed, then any bias from misclassification of exposure will be non-differential and will tend to produce “false-negative” findings. Thus, if positive findings do occur, one can be confident that these are not due to inaccuracies in the exposure data; however, if no association (or only a weak association) is found between exposure and disease, then the possibility of non-differential information bias should be considered. In general, the aim of exposure assessment (see also chapter ▶Exposure Assessment of this handbook) is to (1) ensure that the exposure data is of equal quality in the groups being compared and (2) ensure that the data is of the best possible quality given the former restriction.

Methods of exposure measurement include personal interviews or self-administered questionnaires, diaries, observation, routine records, physical or chemical measurements on the environment, or physical or chemical measurements on the person. Measurements on the person can relate either to exogenous exposure (e.g., airborne dust) or internal dose (e.g., plasma cotinine as a biomarker of tobacco smoke); the other measurement options (e.g., questionnaires) all relate to exogenous exposures.

Traditionally, exposure to many non-biological risk factors (e.g., cigarette smoking) has been measured with questionnaires, and this approach has a long history of successful use in epidemiology. Questionnaires may be self-administered (e.g., postal questionnaires) or interviewer-administered (e.g., in telephone or face-to-face interviews) and may be completed by the study subject or by a proxy (e.g., parental completion of questionnaires in a childhood asthma study). The validity of questionnaire data also depends on the structure, format, content, and wording of questionnaires, as well as methods of administration and selection and training of interviewers. Questionnaires may be combined with environmental exposure measurements (e.g., pollen counts, industrial hygiene surveys) to obtain a quantitative estimate of individual exposures. Questionnaires and environmental measurements have good validity and reproducibility with regard to current exposures and are likely to be superior to biological markers with respect to historical exposures (as described in the following).

Exposure can also be measured using molecular markers of internal dose. However, there are a number of major limitations of many available biomarkers of exposure (Armstrong et al. 1992), particularly with regard to historical exposures (Pearce et al. 1995) with even the best markers of exposures usually reflecting only the last few weeks or months of exposure. Thus, if the aim is to measure historical exposures, then historical information on exposure surrogates may be more valid than direct measurements of current exposure or dose levels. This situation has long been recognized in occupational epidemiology, where the use of work history records in combination with a job-exposure matrix (based on historical exposure measurements of work areas rather than individuals) is usually considered to be more valid than current exposure measurements (whether based on environmental measurements or biomarkers) if the aim is to estimate historical exposure levels (Checkoway et al. 2004). However, some biomarkers have potential value in validation of questionnaires which can then be used to estimate historical exposures. Furthermore, biomarkers of internal dose may have relatively good validity in studies involving an acute effect of exposure such as the triggering of specific asthma attacks.

A more fundamental problem of measuring internal dose with a biomarker is that it is not always clear whether one is measuring the exposure, the biological effect, or some stage of the disease process itself. Thus the findings may be uninterpretable in terms of the causal association between exposure and disease.

A further major problem with the use of biomarkers is that the resulting expense and complexity may drastically reduce the study size, even in a case-control study, and therefore greatly reduce the statistical power for detecting an association between exposure and disease.

Thus, questionnaires and environmental measurements will continue to play a major role in exposure assessment in asthma epidemiology, but biomarkers may be expected to become increasingly useful over time, as new techniques are developed. The emphasis should be on using “appropriate technology” to obtain the most practical and valid estimate of the etiologically relevant exposure.

The appropriate approach (questionnaires, environmental measurements, or biological measurements) will vary from study to study and from exposure to exposure within the same study (or within the same complex chemical mixture, e.g., in tobacco smoke).

56.4.3 Measuring Causal Mechanisms

The causal mechanisms of asthma are still poorly understood. The focus, until recently, was on atopic immune responses, and epidemiological studies therefore often involved measuring atopy. With the introduction of new techniques such as exhaled NO measurements, sputum induction testing, and exhaled breath condensate measurements, epidemiologists now have a wider range of tools available allowing both atopic and non-atopic immunological mechanisms to be studied more fully in population surveys (rather than in laboratory animals as has traditionally been the case).

56.4.3.1 Atopy

Atopy (as defined above) is both an associated condition and a risk factor for developing *allergic* asthma. In the latter context, atopy can often also be considered as an intermediate factor in the causal pathway leading from allergen exposure to allergic asthma. For example, it might be considered inappropriate to control for atopy in a study of dust mite allergen exposure in infancy and asthma at age 5 years, since the development of atopic sensitization might be considered to be an intermediate stage of the causal process leading from dust mite allergen exposure to asthma symptoms. However, atopy may also modify the effect of dust mite allergen exposure, that is, non-asthmatic atopic persons may be more likely to develop asthma symptoms than non-atopic persons at the same level of dust mite exposure. Also, atopics may be more sensitive to irritant exposures causing non-allergic asthma. Thus atopy may be an intermediate factor and/or a modifier of the effects of other exposures.

A positive response to the application of a specific allergen in skin prick testing reflects the production of specific IgE antibodies to the allergen. Skin prick testing therefore provides a convenient test for atopy in epidemiological studies. Although asthma and atopy are often strongly associated, they also occur independently of each other, and "only" one half of all asthma is attributable to atopy (see Sect. 56.3.1) (Pearce et al. 1999). Serum IgE measurements are another well-accepted and widely used measure to assess atopy or atopic sensitization and have the advantage that a larger panel of allergens can be tested. However, they require venipuncture, which may reduce response proportions. Both skin prick tests and serum IgE measurements provide additional information about the potential immunological mechanisms (atopic versus non-atopic), and in case of atopic asthma, they may provide an indication of the potential causal allergen involved.

56.4.3.2 Exhaled NO

Airway inflammation is often considered a hallmark of asthma, but it is difficult to measure, particularly on the scale required for large population studies. Bronchial biopsy, bronchoalveolar lavage, and induced sputum (see below) can be used, but these methodologies are invasive (with exception of suptum induction), time-consuming, and require highly specialized staff. The measurement of the fraction of nitric oxide in exhaled air (FE_{NO}) has been increasingly recognized as a convenient and cost-effective means of non-invasively assessing atopic airway inflammation in asthma, and several instruments are now available utilizing either chemiluminescent or electrochemical NO measurement technologies.

Increased FE_{NO} has been shown to be associated with disease severity, atopy, airway eosinophilia, and bronchial hyperresponsiveness in asthmatics (Alving et al. 1993; Jatakanon et al. 1998; Berry et al. 2005), and corticosteroid therapy has been shown to decrease FE_{NO} (Kharitonov et al. 1996). More recently, it has been suggested that FE_{NO} measurements can be used to optimize corticosteroid dose without negatively impacting upon asthma control (Pijnenburg et al. 2005). Although measuring FE_{NO} shows promise in the diagnosis and management of asthma, other factors not directly related to asthma such as smoking, atopy, height, and sex affect FE_{NO} levels (Kharitonov et al. 1995; Olin et al. 2006; Taylor et al. 2007; Dressel et al. 2008). Furthermore, increased FE_{NO} levels may only be associated with atopic or eosinophilic asthma, as patients with non-atopic or non-eosinophilic asthma appear to have normal FE_{NO} levels, regardless of disease severity (Porsbjerg et al. 2009). Thus, despite FE_{NO} being used in large population studies, it is a poor single diagnostic marker of asthma as has been demonstrated in community-based population studies (Travers et al. 2007). Nonetheless, FE_{NO} may be used to assess allergic or atopic airway inflammation and may therefore provide additional information on the mechanism underlying asthma.

56.4.3.3 Sputum Induction Testing and Exhaled Breath Condensate

Objective monitoring of a wide range of inflammation markers and mediators in asthmatic subjects can provide important information with regard to underlying mechanisms causing asthma symptoms. Exhaled NO (see above) measures only one aspect of airway inflammation which therefore does not provide a complete picture of the inflammatory responses underlying respiratory allergies and asthma. Initial studies on airway inflammation in asthmatics have relied upon bronchoalveolar lavage taken during bronchoscopy, but this technique is invasive and unsuitable for large-scale epidemiological studies. Sputum induction testing and exhaled breath condensate testing are methods involving non-invasive procedures that have the potential to be used in population-based studies.

Sputum induction involves subjects inhaling a 4.5% saline solution at increasing intervals typically starting at 30s to a total of up to 10 to 12 min (Gibson et al. 2000; Simpson et al. 2006). Lung function is measured in between each inhalation period and subjects given inhaled β_2-beta agonist if their FEV_1 drops below 15 to 20% of baseline. The subjects are asked to cough up any sputum after each dose of hypertonic saline and samples will be used to analyze cell types and inflammatory markers and mediators (e.g., IL4, IL5, IL8, ECP). These tests can be done in children from age 8 to 10 years and adults but given the laborious procedures involved the population sample size will necessarily be limited. An increasing number of population-based studies are being conducted using sputum induction testing, and these and previous studies have clearly shown that asthma consists of at least two inflammatory phenotypes (see Sect. 56.3.1). Further studies are currently in progress to assess the clinical and epidemiological importance of these findings.

Exhaled breath condensate (EBC) sampling is a technique that has been used to identify inflammatory markers and mediators in a number of respiratory conditions, including asthma, COPD, acute respiratory distress syndrome (ARDS), and cystic fibrosis (Hoffmeyer et al. 2009). EBC collection involves normal tidal breathing into a tube that is cooled by ice, the breath condensing against the tube's inner surface forming a condensate which can then be poured into vials and frozen for future analysis. Breath condensate volume is mainly determined by duration of ventilation, and it takes up to 10 to 20min to obtain 1 to 3mL of condensate which is sufficient for most analyses.

Most of the condensate consists of water vapor, including a number of volatiles that may serve as markers of airway inflammation. In addition, non-volatile substances (i.e., proteins including a number of cytokines) in the lower respiratory tract can be transported in the form of aerosols in exhaled breath. Successful collection of EBC has been reported using a variety of devices including teflon-lined tubing in an ice bath and specially designed double-wall glass condenser systems (Hoffmeyer et al. 2009). EBC has been shown to contain a range of inflammatory markers and mediators, including nitric oxide, hydrogen peroxide, nitrites, eicosanoids, and various cytokines (Kharitonov and Barnes 2006; Hoffmeyer et al. 2009). The significance of breath condensate analyses is that this technique is straightforward, non-invasive, and completely safe and could be widely used in population-based surveys of asthma (including young children in whom more invasive procedures are not possible). However, EBC results have been highly variable between and within laboratories, and further validation work is therefore required before the technique can "routinely" be used in epidemiological studies of respiratory allergies and asthma (Czebe et al. 2008).

56.5 The Global Burden of Respiratory Allergies and Asthma

56.5.1 Time Trends

It has long been suspected that the prevalence of asthma has been increasing not only in industrialized countries but also in developing countries (Pearce et al. 2000b). However, this has been a particularly difficult issue to resolve because of the lack of systematic standardized studies measuring changes in asthma prevalence over time, and some reviewers have argued that the increases in reported prevalence are largely due to increased awareness, labeling and diagnosis of asthma symptoms (Magnus and Jaakkola 1997). Nevertheless, most studies, which have determined the prevalence of asthma symptoms using the same methodology in the same community at different times, have reported that asthma prevalence has increased in recent decades and that the magnitude of the increase has in some cases been substantial (Table 56.2). Although methodological differences in these studies make it difficult to compare the magnitude of the differences in asthma prevalence between countries, the trend of increasing prevalence among populations in countries of widely differing lifestyles, and ethnic groups is generally consistent.

One of the most informative studies to date is that of Haahtela et al. (1990) who analyzed the medical examination reports of approximately 900,000 conscripts to the Finnish defence forces during 1966–1989, and a proportion of those examined in 1926–1961. During 1926–1961, the prevalence of asthma recorded at call up examinations was in the range of 0.02–0.08%. However, asthma prevalence increased from 0.29% in 1966 to 1.79% in 1989. The authors concluded that the increase was unlikely to be due to improved diagnostic methods and that much of the increase was likely to be real. This conclusion was strengthened by a concomitant rise (from 0.12% in 1966 to 0.75% in 1989) in exemptions and discharges due to asthma. This study is consistent with other evidence that the increases in asthma prevalence in industrialized countries appear to have commenced after the Second World War, particularly in the 1960s and 1970s.

Thus, until recently, most studies had reported that asthma prevalence has increased in recent decades and that the magnitude of the increase had in some cases been substantial.

However, several recent studies have reported either no increase or even a decrease in asthma prevalence over the last ten years. For instance, Bollag et al. (2005) examined time trends in consultations for asthma in primary care in Switzerland and found that overall consultation rates for asthma increased from 1989 to 1994, then stabilized and have declined since 2000. The observation that asthma incidence might be falling is in agreement with several other studies that showed similar time trends for asthma and hay fever (Pearce and Douwes 2005). The best indication of what is currently happening globally is provided by the Phase III of the International Study on Asthma and Allergies in Children (ISAAC); the results of which are discussed below.

The causes of the international time trends in the prevalence of asthma are unclear and are currently a major focus for asthma epidemiology worldwide. An important component of this research process involves standardized international prevalence comparisons (Pearce et al. 1998a). The key problem is to gain information on large numbers of people in random samples collected in a comparable manner across social groups, regions, and countries. Thus, comparisons of asthma prevalence are increasingly being based on a simple comparison of symptom prevalence in a questionnaire survey in a large number of people, followed by more intensive testing of factors related to asthma (e.g., BHR) and risk factors for asthma (skin prick test positivity, serum IgE, and environmental exposures) in a subsample, and a repeat of the prevalence survey over time. This approach has been used in the international

Table 56.2 Changes in asthma prevalence in children and young adults

		Asthma prevalence		
Country	Period	1st study	2nd study	Reference
Australia	1964–1990	19.1%	46.0%	Robertson et al. (1991)
	1982–1992	10.4%	28.6%	Robertson et al. (2004)
	1992–2002	28.6%	23.7%	
	1987–1992	5.6%	9.3%	Campbell et al. (1992)
	1992–1995	9.3%	11.4%	Adams et al. (1997)
	1993–2002	27.2%	20.2%	Robertson et al. (2004)
Canada	1980–1983	3.8%	6.5%	Infante-Rivard et al. (1987)
	1980–1990[c]	140/10,000	256/10,000[a]	Manfreda et al. (1993)
		125/10,000	254/10,000[b]	
England	1956–1975	1.8%	6.3%	Morrison Smith (1976)
	1966–1990	18.3%	21.8%	Whincup et al. (1993)
	1973–1986	2.4%	3.6%	Burney et al. (1990)
	1978–1991	11.1%	12.8%	Anderson et al. (1994)
England and Wales	1970–1981	11.6%	20.5%[a]	Fleming and Crombie (1987)
		8.8%	15.9%[b]	
Finland	1961–1986	0.1%	1.8%	Haahtela et al. (1990)
France	1968–1982	3.3%	5.4%	Perdrizet et al. (1987)
Germany	1991/2–1995/6	3.7%	4.1%	von Mutius et al. (1998)
Israel	1986–1990	7.9%	9.6%	Auerbach et al. (1993)
Italy	1983–1993/5	2.9%	4.4%	Ciprandi et al. (1996)
Japan	1982–1992	3.3%	4.6%	Nishima (1993)
Netherlands	1989–1993	13.4%	13.3%	Mommers et al. (2005)
	1993–1997	13.3%	11.9%	
	1997–2001	11.9%	9.1%	
New Zealand	1969–1982	7.1%	13.5%	Mitchell (1983)
	1975–1989	26.2%	34.0%	Shaw et al. (1990)
Papua New Guinea	1973–1984	0.0%	0.6%	Dowse et al. (1985)
Scotland	1964–1989	10.4%	19.8%	Ninan and Russell (1992)
	1989–1994	19.8%	25.4%	Omran and Russell (1996)
Spain	1994–2003	9.3%	9.3%	Garcia-Marcos et al. (2004)
Sweden	1971–1981	1.9%	2.8%	Aberg (1989)
	1979–1999	2.5%	5.7%	Aberg (1989)
Tahiti	1979–1984	11.5%	14.3%	Liard et al. (1988)
Taiwan	1974–1985	1.3%	5.1%	Hsieh and Shen (1991)
United Kingdom	1991–1998	33.9%	27.5%	Anderson et al. (2004)
USA	1964–1983[d]	183/100,000	284/100,000	Yunginger et al. (1992)
	1971–1976	4.8%	7.6%	Gergen et al. (1988)
	1981–1988	3.1%	4.3%	Weitzman et al. (1992)
	1983–1992	9.2%	15.9%	Farber et al. (1997)
Wales	1973–1988	4.0%	9.0%	Burr et al. (1989)

[a] Men

[b] Women

[c] Prevalence per 10,000 subjects

[d] Incidence rates per 100,000 subjects

survey of asthma prevalence in adults (Burney et al. 1994) and in the International Study of Asthma and Allergies in Childhood (Pearce et al. 1993; Asher et al. 1995; Ellwood et al. 2005).

56.5.2 The European Community Respiratory Health Survey (ECRHS)

In the European Community Respiratory Health Survey (ECRHS), a representative sample of 3,000 adults in each center, aged 20 to 44 years, completed a Phase I screening questionnaire seeking information on asthma symptoms and medication use (Burney et al. 1994). Individuals answering “yes” to waking with an attack of shortness of breath, an attack of asthma, or current asthma medications were defined as “asthmatic.” A random subsample of 600 subjects and an additional sample of up to 150 “asthmatic” individuals in each center were then studied in more detail in Phase II, with measurements of skin prick tests to common allergens, serum total, and specific IgE and bronchial responsiveness to inhaled methacholine, as well as an additional questionnaire on asthma symptoms and medical history, occupation and social status, smoking, the home environment, and the use of medications and medical services. The Phase I results (Burney et al. 1996) included data from 48 centers, predominantly in Western Europe, with only 9 centers from 6 countries (Algeria, Iceland, India, New Zealand, Australia, USA) being from outside of Europe. Phase II was conducted in 37 centers in 16 countries (Burney et al. 1996).

56.5.3 The International Study of Asthma and Allergies in Childhood (ISAAC)

The International Study of Asthma and Allergies in Childhood (ISAAC) (Asher et al. 1995; Ellwood et al. 2005) had a similar study design compared to that of the ECRHS study, with a simple Phase I survey and a more in-depth Phase II survey. However, in order to obtain the maximum possible participation across the world, Phase I (which was conducted in 155 centers in 56 countries) was separated from Phase II (which was conducted in a smaller number of centers), and the Phase I questionnaire modules were designed to be simple and inexpensive to administer. In addition, a video presentation of clinical signs and symptoms of asthma was developed (Shaw et al. 1995) in order to minimize translation problems. The population of interest was school children aged 6 to 7 years and 13 to 14 years within specified geographical areas. The Phase I findings, involving more than 700,000 children, showed striking international differences in asthma symptom prevalence (Asher et al. 1998; Beasley et al. 1998). Figure 56.1 shows the international patterns of 12-month period prevalence of wheezing in 13 to 14-year-olds (based on the question “have you had wheezing or whistling in the chest *in the last 12 months*?”).

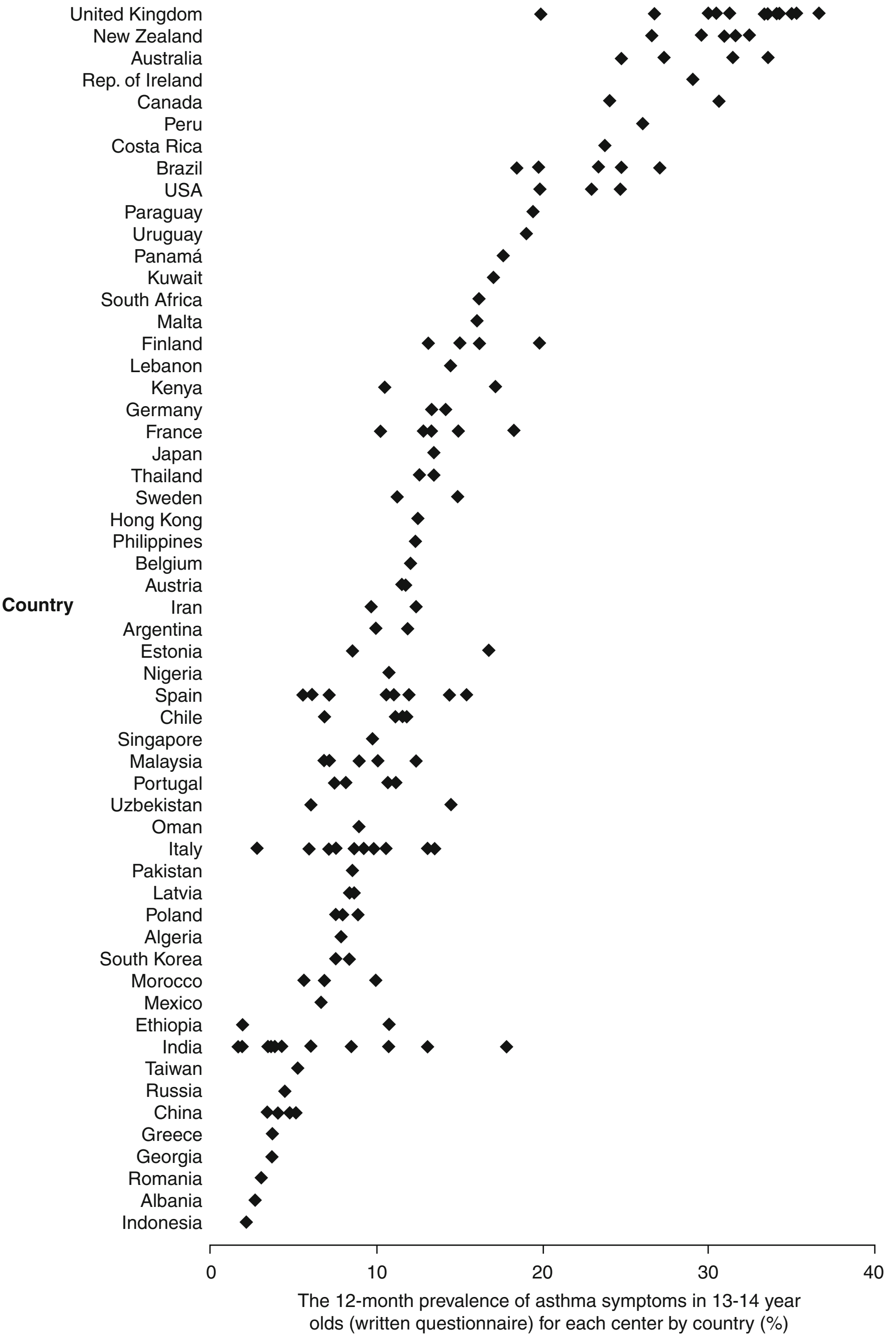


Fig. 56.1 Wheeze in the previous 12 months for each center by country ordered according to the mean prevalence for all centers in the country (Source: Beasley et al. 1998)

ISAAC Phase II was conducted in 30 centers in 22 countries and involved parental questionnaires ($n = 54{,}439$), skin prick tests ($n = 31{,}759$), and serum IgE measurements ($n = 8{,}951$). House dust samples to measure indoor allergens were also collected (Weinmayr et al. 2007).

Phase III involved a repeat of the Phase I survey after 5 to 10 years in 106 centers in 56 countries in children aged 13 to 14 years ($n = 304{,}679$) and in 66 centers in 37 countries in children aged 6 to 7 years ($n = 193{,}404$) (Asher et al. 2006; Pearce et al. 2007).

56.5.4 What Do the ECRHS and ISAAC Studies Show?

The ISAAC and ECRHS studies provide, for the first time, a picture of global patterns of asthma prevalence and identify the key phenomena which future research must address and attempt to explain.

Firstly, both studies show a particularly high prevalence of reported asthma symptoms in English-speaking countries (Fig. 56.1), that is, the British Isles, New Zealand, Australia, the United States, and Canada (Burney et al. 1996; Asher et al. 1998). This appears to be unlikely to be entirely due to translation problems, since the same pattern was observed with the ISAAC video questionnaire (Asher et al. 1998).

Secondly, the ISAAC survey showed that centers in Latin America also had particularly high symptom prevalence (Fig. 56.1). This finding is of particular interest in that the Spanish-speaking centers of Latin America showed higher prevalences than Spain itself, in contrast to the general tendency for more affluent countries to have higher prevalence.

Thirdly, among the non-English-speaking European countries, both studies show high asthma prevalence in Western Europe, with lower prevalences in Eastern and Southern Europe. For example, in the ISAAC survey, there is a clear northwest-southeast gradient within Europe, with the highest prevalence in the world being in the United Kingdom and some of the lowest prevalences in Albania and Greece (Asher et al. 1998). The west-east gradient was particularly strong; in particular, there was a significantly lower prevalence in the former East Germany than in the former West Germany.

Fourthly, Africa and Asia generally showed relatively low asthma prevalence (Fig. 56.1). In particular, prevalence was low in developing countries such as China and Indonesia, whereas more affluent Asian countries such as Singapore and Japan showed relatively high asthma prevalence. Perhaps the most striking contrast is between Hong Kong and Guangzhou which are close geographically and involve the same language and predominant ethnic group; Hong Kong (the more affluent city) had a 12-month period prevalence of wheeze of 10.1%, compared with 2.0% in Guangzhou (the less affluent city).

Fifthly, in contrast to the asthma findings, the highest prevalences of rhinitis symptoms were reported from centers scattered throughout most regions of the world, including Western Europe, Africa, North America, and Southeast Asia; the highest prevalences of eczema were generally in centers of high latitude, including Scandinavia and New Zealand, although there were some notable exceptions including some centers in South America and Africa (Ethiopia). Thus, although the prevalences of these conditions were correlated, the association was not particularly strong, and there were numerous centers which had high prevalence for asthma but not for rhinitis and/or eczema, and vice versa, suggesting that the major risk factors are different for these related disorders or that they involve different latency periods and time trends.

Sixthly, the ISAAC Phase II study showed that the link between atopic sensitization and asthma symptoms differed strongly between populations and increased with economic development (Weinmayr et al. 2007); the association between atopy and flexural eczema was also weak and positively linked to gross national income (Flohr et al. 2008).

Finally, asthma prevalence has peaked or even begun to decline in many affluent countries, whereas asthma symptom prevalence continues to rise in less affluent countries. In particular, ISAAC Phase III showed that international differences in asthma symptom prevalence have reduced, particularly in 13- to 14-year-olds, with decreases in prevalence in English-speaking countries and Western Europe and increases in prevalence in regions where prevalence was previously low including Africa, Latin America, and parts of Asia (Fig. 56.2) (Asher et al. 2006; Pearce et al. 2007). Similarly, Phase II of the European Respiratory Health Survey (ECRHS) found no further increase in current or severe asthma symptoms (Chinn et al. 2004). Nonetheless, a significant increase in diagnosed asthma was observed which most likely reflects changes in diagnostic labeling and/or medical treatment for mild and/or moderate asthma (Weiland and Pearce 2004). The asthma symptom prevalence increases in Africa, Latin America, and parts of Asia indicating that the global burden of asthma is continuing to rise, but the global prevalence differences are lessening.

56.6 Causes of Respiratory Allergies and Asthma

56.6.1 Risk Factors

56.6.1.1 Genetic Factors

Asthma is multifactorial in origin and influenced by multiple genes and environmental factors. Thus, it is not inherited in the simple Mendelian fashion that is characteristic of single-gene disorders. A particular genetic factor may affect one or more aspects of the complex etiological processes potentially involved in asthma including atopic sensitization, bronchial hyperresponsiveness (BHR), airway inflammation, innate immunity, etc. Whether this genetic potential is expressed will depend on various factors, including whether sufficient exposure to environmental factors occurs. Investigating possible genes for the individual etiological factors is also fraught with difficulties, since control of these factors (e.g., IgE production and BHR) are also multifactorial (Zamel et al. 1996).

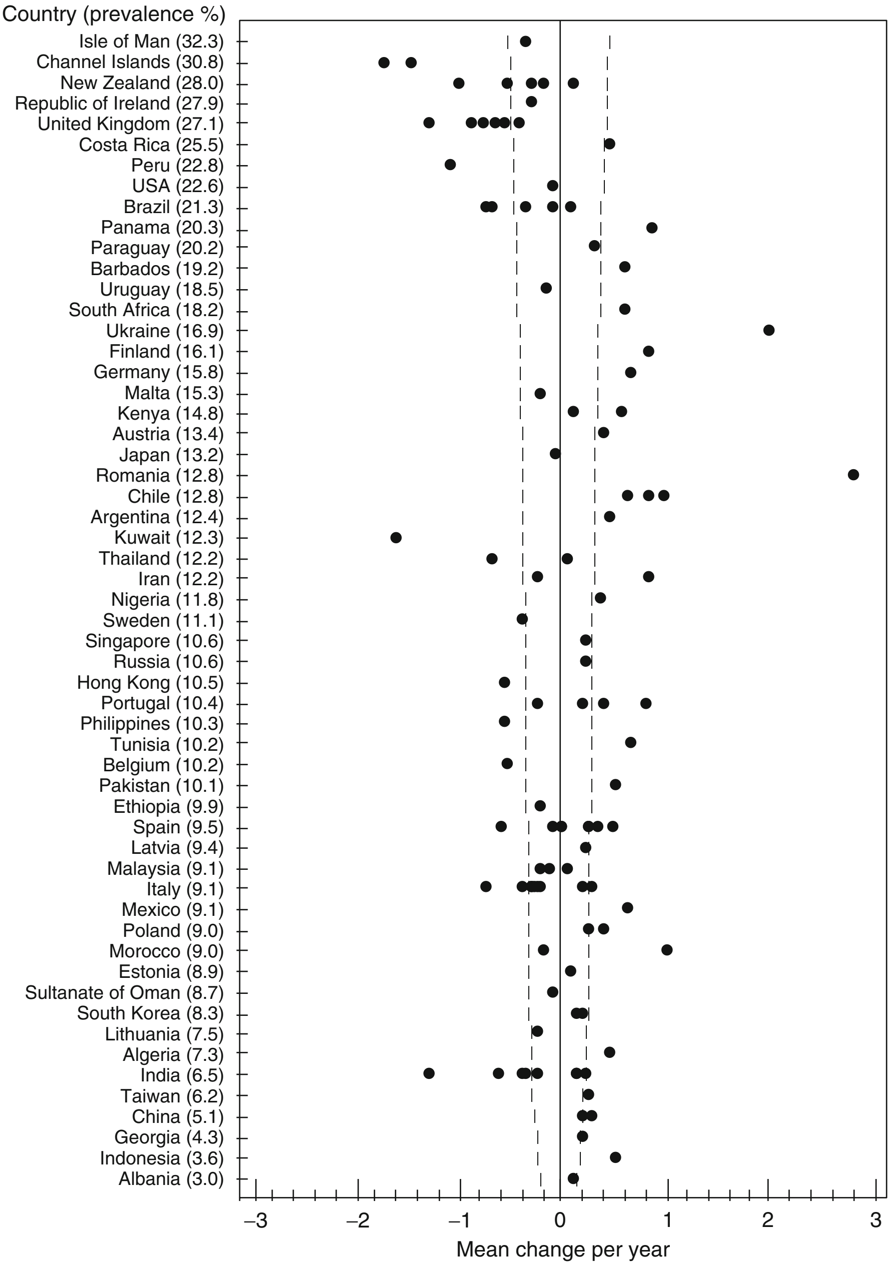


Fig. 56.2 Ranking plot showing the change per year in prevalence of current wheeze (wheeze in the last 12 months) in 13- to 14-year-old children for each center by country, with countries ordered by their average prevalence (for all centers combined) across Phase I and Phase III [the plot also shows the confidence interval about zero change for a given level of prevalence (i.e., the average prevalence across Phase I and Phase III) given a sample size of 3,000 and no cluster-sampling effect] (Source: Pearce et al. 2007)

Also, as noted earlier, asthma is an extremely heterogeneous disease including a variety of phenotypical and clinical manifestations which are likely to be associated with different (combinations of) genes. Another potential source of phenotypic variability is that asthma development, exacerbations, and progression may involve different environmental triggers and genetic factors.

Nonetheless, it is well established that people with a family history of asthma are more likely to develop asthma themselves, and parental asthma is a stronger predictor of asthma in the offspring than parental atopy. However, this association is not necessarily due only to genetic factors but could also reflect similar lifestyles and exposures in family members (Sandford et al. 1996).

Some indication of the possible contribution of genetic factors in asthma is given by studies of twins. For example, Edfors-Lubs (1971) analyzed data on 7,000 twin pairs from the Swedish Twin Registry and found that concordance of asthma in monozygotic twins was greater than in dizygotic twins. However, the concordance was still only 19%, and even this may in part be due to similar environmental exposures in monozygotic twins, including a common intrauterine environment (Godfrey et al. 1994). Other large population studies have yielded similar findings, but this may be because these studies have determined asthma on the basis of questionnaires or hospital and pharmacy records, whereas smaller studies with more intensive diagnostic methods have generally yielded higher concordance (Sandford et al. 1996).

Recent genome-wide association studies (GWAS) including two meta-analyses (GABRIEL and EVE consortia; Moffatt et al. 2010; Ober and Yao 2011) have shown a number of loci that were consistently associated with asthma. However, these associations were relatively weak and the genetic variants very common. As a result, despite these associations being highly statistically significant, the sensitivity and specificity for identifying asthmatics based on these genetic variants was poor (Ober and Yao 2011). Also, many "asthma genes" previously indentified in candidate gene association studies (using a hypothesis-driven approach) were not replicated in these large GWAS meta-analyses. This may be because GWAS studies are not able to detect some of the rarer variants (Michel et al. 2010), or, alternatively, the previously indentified candidates may be false-positive results. Whatever the reason, it is clear that our knowledge of the genetics of asthma is currently still very limited, but future GWAS for asthma on well-defined asthma phenotypes may shed some more light on the importance of genetics in the development of asthma.

56.6.1.2 Demographic Factors

There are a variety of demographic factors which are associated with asthma including age (Anderson et al. 1992), sex (Anderson et al. 1992), and ethnicity (Pattemore et al. 2004). Age is the demographic factor which is most strongly related to asthma symptom prevalence, with symptoms usually declining at or before the onset of puberty (Kimbell-Dunn et al. 1999).

Asthma incidence and prevalence are consistently lower in females than in males before age 12 years, whereas during adolescence and adulthood there is evidence of higher incidence and prevalence in females (Kimbell-Dunn et al. 1999). One possible explanation is that the average age of onset in childhood and adolescence may be later in females. Levels of cord blood IgE are lower at birth in girls than in boys (Weeke 1992), indicating a lower risk of the subsequent development of asthma. Some authors have noted that boys have smaller airways than girls, relative to lung size, and that this may explain the greater frequency and severity of lower respiratory tract illness in boys, even though infection rates are similar for both sexes (Martinez et al. 1988). Alternatively, it is possible that boys have more exposure to factors that increase asthma incidence or duration. The relatively higher prevalence (or smaller reduction in prevalence) in females than in males after puberty could be due to hormonal influences on allergic predisposition, airway size, inflammation, and smooth muscle vascular functions (Redline and Gold 1994). Premenstrual asthma may be especially relevant to the hormonal involvement of asthma since it may not only cause asthma exacerbations but may thereby affect the frequency and duration of asthma symptoms, resulting in an increase in the prevalence of "current asthma."

Studies in the 1960s and 1970s suggested that asthma is more common in children in the higher social classes. There has been less evidence of social class differences as the diagnosis of asthma has become more widespread (Littlejohns and Macdonald 1993), even though diagnostic labeling of wheezing in adults differs by social class (Littlejohns et al. 1989). However, severe asthma appears to be more common in children in the lower social classes (Stewart et al. 2001) and in some disadvantaged ethnic groups (Pattemore et al. 2004), and low socioeconomic status is associated with hospital admissions for asthma (Watson et al. 1996) and with reduced lung function in adults. This could represent either a greater prevalence of asthma in disadvantaged groups, or increased severity due to environmental factors (e.g., environmental tobacco smoke, nutrition, occupational exposures) (Eagan et al. 2002; Ellison-Loschmann et al. 2007), or inadequate disease management and poor access to health care (Ellison-Loschmann and Pearce 2006).

56.6.1.3 Obesity

The specific mechanisms linking body weight and asthma are unclear, but several have been proposed including (1) common etiologies, (2) comorbidities, (3) mechanical factors, and (4) adipokines, that is, cytokines secreted by adipose tissue (Shore 2007). Like asthma, the prevalence of overweight and obesity has increased dramatically in the past few decades in many regions of the world (Burney 2002). Studies have shown associations between body weight and asthma in both adults (Braback et al. 2005) and children (von Mutius et al. 2001). Prospective studies of children suggest that obesity precedes asthma signaling a causal link (Gilliland et al. 2003; Gold et al. 2003), as do studies showing associations between asthma and both weight gain and weight loss in adults (Hakala et al. 2000). Nonetheless, some studies have failed to show an association (Brenner et al. 2001), while others showed an association only in one sex (Mannino et al. 2006). Several reported an association only with respiratory symptoms (e.g., wheeze), but not BHR (Bustos et al. 2005), although this may merely indicate that obesity increases asthma risk through mechanisms other than BHR. Obesity may also increase severity in subjects with preexisting asthma (Akerman et al. 2004).

56.6.1.4 Diet

Many studies have investigated the effects of breast-feeding on allergies and asthma with some studies showing protective effects, some showing no effect, and others suggesting that breast-feeding is a risk factor (Friedman and Zeiger 2005). A recent longitudinal study found that exclusive breast-feeding was associated with a slightly reduced risk of asthma and atopy at age 7, but an increased risk at age 14 and 44 (Matheson et al. 2007). Another infant cohort study showed that breast-feeding did not protect against atopy and asthma but even increased the risk at age 9 to 26 years (Sears et al. 2002). A cluster randomized trial reported similar findings for allergies and asthma at age 6.5 years (Kramer et al. 2007).

Other nutritional factors may also play a role in the etiology of asthma. In particular, it has been speculated that the increase in asthma prevalence may be due to a change in dietary patterns in the past few decades, that is, as cultures have become more “westernized,” they have shifted from a tradition of growing and consuming locally grown foods to consuming more processed foods with an overall increase in the intake of refined sugars, fats, and additives, as well as a reduction in the intake of fresh fruits, vegetables, and fish. Several observational studies have shown protective effects of fruit and vegetables, whole grain products, and fish (Devereux 2007), and these findings are consistent with an ecological analysis of the ISAAC Phase I survey (Ellwood et al. 2001). Fruit, vegetables, and wholegrain products are rich in antioxidants and may reduce airway inflammation by protecting the airways against both endogenous and exogenous oxidants (Devereux 2007). Fish oils are rich in n-3 polyunsaturated fatty acids which may also protect against airway inflammation and subsequent symptoms of asthma (Devereux 2007). However, dietary supplement studies focusing on antioxidants and n-3 fatty acids have not shown convincing evidence of a protective effect on allergies and asthma (Reisman et al. 2006; Almqvist et al. 2007).

56.6.1.5 Outdoor Air Pollution

The role of outdoor air pollutants (particulate matter, ozone, nitrogen dioxide, and sulfur dioxide) in asthma and other diseases has been extensively studied and debated. An association between measures of distance to major roads or traffic density and asthma symptoms has been found in a number of European countries (WHO 2005). Also, a large number of studies have reported associations between direct measurements of air pollution levels and exacerbation of preexisting asthma, both in children and adults (Boezen et al. 1999; WHO 2005). Some studies, including a recent birth cohort study (Brauer et al. 2007), have also suggested that air pollution may cause *new onset* of asthma and allergic disease. In particular, several large prospective studies have suggested a role for ozone (McDonnell et al. 1999; McConnell et al. 2002), although significant associations with some asthma outcomes were also shown for PM2.5, soot, and NO_2 (Brauer et al. 2007). Nonetheless, although it is clear that air pollution can provoke exacerbations in preexisting asthma and a positive association between outdoor air pollution and asthma prevalence at the population level has been shown (Asher et al. 1998), the weight of evidence does not currently support a *major* role for outdoor air pollution as a cause of the initial development of asthma.

56.6.1.6 Indoor Air Pollution

Little is currently known about the contribution of indoor air pollutants (other than environmental tobacco smoke) to the incidence and prevalence of asthma. The range of potential pollutants is large, the determinants of ambient levels involve a complex interaction of lifestyle and building factors, and precise measurement of airborne concentrations is difficult. Nitrogen dioxide from burning fossil fuels has received by far the most attention, while sulfur dioxide from burning sulfur-containing coal or gas, mosquito coil smoke, and formaldehyde from wood preparation have also been considered. Particulates from open or closed wood and coal burning fires have received less attention in developed countries but have been studied in developing countries where very high indoor levels have been encountered.

Damp indoor environments and indoor fungal exposure may also play a role as demonstrated in a large number of studies conducted across many geographical regions (Douwes and Pearce 2003). However, although it has been concluded that the evidence for a causal association between dampness and respiratory morbidity is strong, it is not clear whether indoor dampness *causes* or “only” *exacerbates* preexisting respiratory conditions such as asthma (Douwes and Pearce 2003).

56.6.1.7 Tobacco

Similarly, the evidence for a role of tobacco smoke in asthma is strongest for increases in severity in children who already have asthma, whereas the evidence for the initial occurrence of asthma (incidence) is less conclusive. In particular, several recent reviews and meta-analyses differ in their conclusions about the role of second-hand tobacco smoke (SHS). The US Environmental Protection Agency (EPA) and Californian EPA concluded that SHS was causally associated with the development of asthma in children (EPA 1992; OEHHA 1997). The 2006 Surgeon General’s report on “health effects from involuntary exposure to tobacco smoke” (DHHS 2006) concluded that the evidence was suggestive, but not sufficient to infer a causal relationship. This analysis was based on a previous meta-analysis conducted by Strachan and Cook (1998) and did not include the most recent epidemiological studies. However, the more recent meta-analysis including studies published between 1970 and 2005 concluded that household SHS exposure was positively and consistently associated with the incidence or new onset asthma (Vork et al. 2007) not only in younger but also older children.

The evidence on active smoking as a risk factor is also conflicting with some studies reporting only exacerbations (Siroux et al. 2000), whereas others have also documented a dose-related risk of new-onset asthma in adolescents and adults (Eagan et al. 2002). Overall, it therefore appears that environmental tobacco smoke is a cause of asthma exacerbations and that in addition it may also be involved in the development of asthma itself.

56.6.1.8 Occupational Exposures

Occupational asthma (OA) is the most common occupational respiratory disease in developed countries. For example, asthma accounted for 28% of cases reported to the United Kingdom surveillance of work-related and occupational respiratory diseases (SWORD) project (Meredith et al. 1991). Estimates of the total proportion of adult asthma which is thought to be occupational in origin range from 2 to 15% in the United States (Chan-Yeung and Malo 1994), 15% in Japan (Chan-Yeung and Malo 1994), 5% in Spain (Kogevinas et al. 1996), 2% to 3% in New Zealand (Fishwick et al. 1997), and 2% to 6% in the United Kingdom (Meredith and Nordman 1996). However, much higher estimates have also been reported. For example, a study of the entire employed population of Finland from 1986 to 1998 estimated the attributable fraction of adult-onset asthma due to occupation to be 29% for men and 17% for women (Karjalainen et al. 2001). More than 250 agents have been identified as causes of OA (Department of Health and Senior Services, State of New Jersey 2006). Some of the most common occupational asthmagens include flour/grain dusts, wood dusts, latex allergens, and isocyanates.

56.6.1.9 Respiratory Viruses

Viral infections are common causes of exacerbations of asthma (Johnston et al. 1995). In fact, respiratory viral infections are detected in the majority of asthma exacerbations (80% to 85% in children and 75% to 80% in adults); of these, about 60% are rhinoviruses (Johnston 2007). There is also a strong association between viral infections and hospital admission for asthma in both children and adults.

Viral infections may also be involved in the development of asthma, but the evidence is less clear. Several long-term longitudinal studies have shown that respiratory syncytial virus (RSV) infections increase the risk of subsequent recurrent wheezing and asthma in early childhood (Stein et al. 1999; Sigurs et al. 2005). However, this risk may progressively decrease with increasing age (Stein et al. 1999). Other viruses have also been associated with asthma development including human rhinovirus (HRV) which may in fact be a more important risk factor than RSV (Lemanske et al. 2005). The mechanisms of viral-induced asthma are poorly understood, but it has been speculated that impaired innate immune responses may play a crucial role (Johnston 2007).

56.6.1.10 Medications

Antibiotics The "hygiene hypothesis" postulates that growing up in a more hygienic environment with less microbial exposure may increase the risk of allergies and allergic asthma (see Sect. 56.6.3). A corollary of the hygiene hypothesis is that antibiotic use may increase the risk of asthma by reducing the protective effect of bacterial infections and/or disruption of the normal gut bacterial flora (Farooqi and Hopkin 1998; Mendall and Kumar 1998). However, the epidemiological evidence of an association between exposure to antibiotics (as well as infection) and the development of asthma has been conflicting (Celedon et al. 2002; Cohet et al. 2004; Kummeling et al. 2007).

One of the first reports was that of Farooqi and Hopkin (1998) who found significant associations between treatment with oral antibiotics in the first two years of life and subsequent asthma, hay fever, and eczema at age 12 to 20 years in Oxfordshire. The association was stronger for infections treated with broad-spectrum antibiotics and increased with the number of antibiotic courses received. McKeever et al. (2002a) found that antibiotic exposure in early life was associated with an increased risk of asthma diagnosis, but the association was reduced when the data were adjusted for consulting behavior. They also reported that exposure to antibiotics in utero was associated with a dose-related increase in the child's risk of asthma, hay fever, and eczema (McKeever et al. 2002b). However, Celedon et al. (2004) found no significant association between antibiotic use in the first year of life and the subsequent development of asthma, allergic rhinitis, or eczema at age 5 years. Wjst et al. (2001) found a dose-dependent association of antibiotic use with asthma diagnosis in children aged 5 to 14 years. However, the authors suggested that their findings may be due to reverse causation.

More recently, Foliaki et al. (2009) analyzed the ISAAC Phase III data (see Sects. 56.5.3 and 56.5.4) and found that the reported use of antibiotics in the first year of life was associated with parental-reported symptoms of asthma in 6 to 7-year-old children, following adjustment for other asthma risk factors. The association was present in all major regions of the world (with the possible exception of Africa). Similar associations were observed between the use of antibiotics in the first year of life and the risk of severe asthma symptoms and "asthma ever." Weaker (but still statistically significant) associations were also observed for symptoms of rhinoconjunctivitis and eczema.

Paracetamol It has been reported that prenatal paracetamol (or acetaminophen) use during pregnancy was a risk factor for asthma, wheezing, and total IgE in the offspring at 6 to 7 years of age (Shaheen et al. 2002, 2005) which was unlikely to be confounded by unmeasured behavioral factors linked to paracetamol use (Shaheen et al. 2010). Similarly, several cross-sectional and longitudinal studies have reported that paracetamol use was associated in a dose-dependent manner with an increase in asthma in children and adults and also new-onset asthma in adults (Shaheen et al. 2000, 2008; Barr et al. 2004). Furthermore, national per capita consumption of acetaminophen was ecologically associated with the prevalence of wheeze, diagnosed asthma, and bronchial hyperresponsiveness in Western Europe (Newson et al. 2000).

In the ISAAC Phase III data (see Sects. 56.5.3 and 56.5.4), the reported use of paracetamol for fever in the first year of life was associated with an increased risk of current asthma symptoms (Beasley et al. 2008). There was a dose-dependent increased risk of current asthma symptoms. However, the association was weaker than that observed for antibiotic use in the same data set (see above; Foliaki et al. 2009), and the association reduced (from 1.77 to 1.46) when adjusted for antibiotic use and other asthma risk factors, indicating that the elevated risk was, at least in part, due to confounding. In particular, it has been suggested that the observed association is due to confounding by indication or reverse causation (Lowe et al. 2009).

However, confounding by indication or reverse causation are unlikely to fully explain the positive associations in birth cohort studies (Shaheen et al. 2002, 2005, 2010), and longitudinal studies in adults that focused on new-onset asthma (Barr et al. 2004). The underlying mechanisms are unclear, but it has been suggested that paracetamol decreases glutathione levels in the lung, which may predispose to oxidative injury, bronchospasm, and an increased Th_2 response (Shaheen et al. 2002). Interestingly the use of paracetamol has increased considerably (replacing aspirin) in the 1970s and 1980s (Varner et al. 1998) suggesting that the increased use of paracetamol may account for some of the increasing prevalence of childhood asthma.

56.6.1.11 Allergens

Indoor allergens, particularly house dust mite allergens, are perhaps the group of possible asthma risk factors that have received the greatest attention. It is well established that in sensitized asthmatics, allergen exposure can trigger asthma attacks and that prolonged exposure can lead to the prolongation and exacerbation of symptoms. However, most studies in children show only weak associations between allergen exposure and current asthma, even when the analyses are restricted to atopic patients and allergen avoidance has been accounted for (Pearce et al. 2000c). Also, secondary intervention trials have had mixed results (Gotzsche et al. 1998).

In fact, although there is good evidence for asthma exacerbations, the evidence for new onset asthma is much weaker (Pearce et al. 2000c). The key study linking allergen exposure in infancy to the subsequent development of asthma is that of Sporik et al. (1990) who followed 67 children with a family history of atopy. They found an association between dust mite allergen levels and mite sensitization and an association between exposure to more than 10ug/g in the first year of life and a history of wheezing, although this association was not statistically significant (odds ratio $(OR) = 2.3$, $\mathrm{p} = 0.17$). There were non-significant associations of dust mite levels with “active wheezing and BHR” ($p = 0.08$) and with “receiving medication” ($p = 0.10$).

More recent longitudinal birth cohort studies have found little or no association between early dust mite allergen exposure and asthma later in childhood (Burr et al. 1993; Corver et al. 2006; Tepas et al. 2006). For example, Burr et al. (1993) conducted a longitudinal study among 453 infants in South Wales with a family history of allergic diseases. Doctor diagnosed asthma and wheezing at age 7 years was neither associated with mite allergen exposure as determined in the first 12 months nor with dust mite levels measured at 7 years of age (odds ratios were not given). Similarly, in the German Multicentre Allergy Study, levels of mite and cat allergens in early life remained strongly related to specific sensitization at age 3 to 7 years, but no dose-response relationship between allergen exposure and any measure of asthma/wheeze at 7 years of age was found (Lau et al. 2000, 2002). Dust mite allergens are therefore unlikely to play a major role in the initial development of asthma.

There are several other indoor and outdoor allergens that have been suggested to be associated with the development of asthma including cat, dog, cockroach, and *Alternaria* allergens. However, the evidence for a causal relationship is even weaker than for house dust mite allergens (Pearce et al. 2000c). In fact, several studies have even reported that having a pet early in life protects against the development of asthma (see Sect. 56.6.3).

56.6.1.12 Emotional Stress

Until the second half of the twentieth century, the predominant view was that asthma was a psychosomatic disorder in which emotional stress was the key factor in its etiology; the condition was therefore commonly referred to as "asthma nervosa" (Salter 1860). With the recognition of causal environmental exposures such as pollen (Blackley 1873) and house dust (Osler 1892) in the latter half of the nineteenth century and the increased understanding of the inflammatory mechanisms underlying asthma in the second half of the twentieth century (Holgate 2004), the notion that asthma was caused by emotional stress slowly lost support. Thus, many asthma researchers currently regard emotional stress as predominantly a consequence of the disease or as an external factor which may trigger exacerbations in those with preexisting asthma. Associations between stress/anxiety and asthma in (cross-sectional) epidemiological studies are therefore generally dismissed as examples of reverse causation. However, several prospective studies in young children and adults suggest that this assumption may not be justified (Douwes et al. 2010, 2011).

A recent birth cohort in 5,810 children aged 7.5 years showed a strong association between prenatal maternal anxiety symptoms (as an indicator of stress during fetal life) and asthma prevalence at 7.5 years (Cookson et al. 2009). Children of mothers in the highest quartile of anxiety scores were 64% more likely to have asthma compared with those in the lowest quartile (*OR* = 1.64, 95% confidence interval (CI) = 1.25–2.17). Postnatal anxiety was not associated with asthma when adjusted for prenatal anxiety, suggesting that the prenatal period may be particularly critical.

The effects of stress/anxiety may not be limited only to early life events. For example, a 13-year follow-up study which assessed the associations between war-related stressors and new onset asthma in 2,066 elderly (50 to 69 years) Kuwaiti civilians following the Iraqi invasion in 1990 and the subsequent 7-month occupation found a dose-response relationship with asthma incidence (assessed as a self-reported doctor diagnosis of asthma), after adjusting for potential confounders including air pollution related to burning oil fires (Wright et al. 2010). The highest stress level more than doubled the risk of new onset asthma (*OR* = 2.3, 95% CI = 1.3–3.9).

The mechanisms underlying the association between stress and asthma remain far from clear. Several studies have suggested that stress acts principally through altered regulation of the hypothalamic-pituitary-adrenal (HPA) axis and sympathetic-adrenomedullary (SAM) nervous system (Vig et al. 2006). The subsequent change in levels of neurohormones (and in particular endogenous glucocorticoids) has considerable immunomodulatory effects, including an atopy (or Th_2)-biased response favoring allergic outcomes (von Hertzen 2002). In addition to neuroimmunomodulatory effects, it is plausible that direct neurogenic mechanisms may underlie at least part of the association between stress and asthma (Miller et al. 2009).

56.6.2 Can the Traditional Risk Factors Explain the International Patterns and Time Trends?

Although there is substantial evidence that various environmental risk factors can increase the risk of developing asthma, there is little evidence that the traditional risk factors can account for the global prevalence increases or the international prevalence patterns that have been observed. The increases in asthma prevalence cannot be due to genetic factors, since they are occurring too rapidly, and the rapidity of the increases indicates that genetic factors alone are unlikely to account for a substantial proportion of asthma cases (Douwes and Pearce 2002), although genetic susceptibility to changing environmental exposures may play an important role.

The global patterns of asthma prevalence are also inconsistent with the hypothesis that air pollution is a major risk factor for the development of asthma (Asher et al. 1998, 2006; Beasley et al. 1998). Regions such as China and Eastern Europe where there are some of the highest levels of traditional air pollution such as particulate matter and SO_2 generally have lower asthma prevalence than the countries of Western Europe, North America, Australia, and New Zealand which have lower levels of pollution. It also appears very unlikely that the international prevalence patterns can be explained by differences in smoking (Mitchell et al. 2002) or in occupational exposures.

Allergen exposure is the risk factor that has perhaps received the most attention as a possible cause of the global increases in prevalence of asthma and allergies. In particular, it has been suggested that increases in indoor allergen exposures, through changes in lifestyle such as wall-to-wall carpeting, cold water washing, greater time spent indoors watching television, etc., could account for the global increases in asthma prevalence (Sporik et al. 1990). However, the only study of English homes at two time points (1979 and 1989) did not find any change in house dust mite allergen levels (Butland et al. 1997), although marked increases have been observed in Australian studies (Peat et al. 1996).

The ISAAC (Asher et al. 1998) and ECRHS studies (Burney et al. 1996) have consistently found uniformly high levels of asthma prevalence in centers in English-speaking countries, even though there is a wide variation in house dust mite levels across these countries (Martinez 1997). In geographical areas in which dust mite exposure is very low or absent, including desert regions and mountainous regions, the prevalence of asthma is as high or even higher than that in other areas where house dust mite exposure is high (Martinez 1997).

Other available evidence on the association between allergen exposure and the subsequent risk of asthma at the population level is also less than persuasive. For example, Leung et al. (1997) reported that asthma prevalence was high in

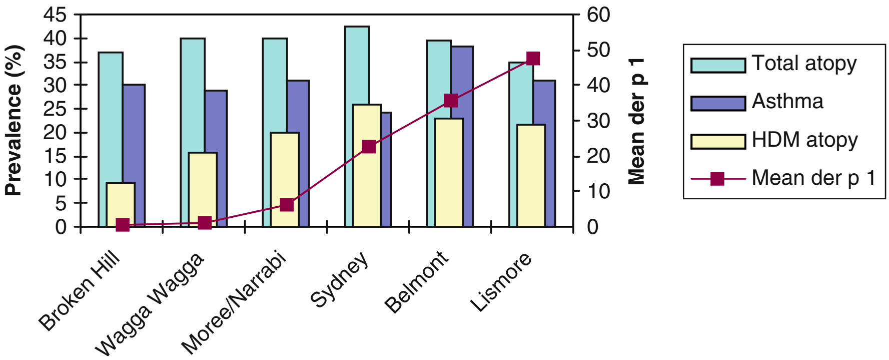


Fig. 56.3 Mean Der p 1 levels and prevalence (%) of house dust mite atopy, total atopy, and asthma in six areas of Australia

Hong Kong (6.6% for asthma ever) and low in San Bu, China (1.6%), but exposures to house dust mite allergen were similar in Hong Kong and San Bu. Similarly, Fig. 56.3 shows data from seven Australian surveys in centers with widely differing levels of mite allergen exposure; the overall prevalences of sensitization and asthma were both unrelated to the levels of house dust mite allergen (Der p 1) exposure in the six centers. The dominant allergen varied between regions, but there was little overall difference in the prevalence of sensitization or of asthma despite the major differences in mite allergen levels. Similarly, von Mutius et al. (1994) found that asthma was significantly higher in Munich, West Germany (5.9%) than in Leipzig, East Germany (3.9%), and this paralleled the pattern of skin prick test positivity (19.2% and 7.3%). However, house dust mite allergen levels were similar in the east and the west (Hirsch et al. 1998).

The other asthma risk factors (e.g., diet, obesity, paracetamol, emotional stress) may significantly contribute to the observed time trends and international patterns of asthma prevalence, but there is little evidence for this currently.

56.6.3 Protective Factors

Recent research has shifted attention from allergens that may cause sensitization and/or provoke asthma attacks, to factors that may "program" the initial susceptibility to asthma, through allergic or non-allergic mechanisms. This also in part involves a shift of attention from risk factors for asthma to protective factors and the possible role of the loss of protective factors in the global increases in asthma prevalence.

56.6.3.1 The Hygiene Hypothesis

The "hygiene hypothesis" postulates that growing up in a more microbiologically hygienic environment may increase the risk of developing respiratory allergies and has been prompted by evidence that overcrowding and unhygienic conditions were associated with a lower prevalence of atopy, eczema, hay fever, and asthma (Strachan 1989). Having a large number of siblings (especially older siblings) and attendance at day care centers were determined to be particularly protective (Ball et al. 2000). An increase in infections has been proposed as an explanation for these findings and several studies have in fact shown a direct association between infections (e.g., Hepatitis A, measles) or immunization with BCG (Bacillus Calmette-Guérin) against tuberculosis and a lower prevalence of atopy and allergies (Shaheen et al. 1996; Matricardi et al. 1997). However, the results for airborne viruses (measles, mumps, rubella, and chickenpox) and BCG vaccination were inconsistent (Alm et al. 1998; Matricardi et al. 2000).

Exposure to specific microbial agents with strong proinflammatory properties, such as bacterial endotoxin, has also been suggested to be protective (Douwes et al. 2002a). Studies in both rural and non-rural environments have reported a significant inverse association between indoor endotoxin levels and atopic sensitization (Gehring et al. 2001), hay fever, and atopic asthma (Braun-Fahrlander et al. 2002). In contrast, a birth cohort study conducted by the same researchers found that early endotoxin exposure was associated with an *increased* risk of atopy at the age of 2 years (Bolte et al. 2003). However, two similar birth cohort studies found a protective effect on atopy in 2-year-olds (Bottcher et al. 2003) and asthma symptoms in 4-year-olds (Douwes et al. 2006).

Thus, the evidence is currently mixed as to whether endotoxin exposures may protect against atopy and allergic asthma. If there is a causal association, most of the evidence points toward endotoxin, but other pathogen-associated molecular patterns (PAMPs) may be equally (or more) important. There is evidence that exposure to peptidoglycans, CpG-containing DNA, and certain viruses may also reduce the risk of atopic disease (Douwes et al. 2004). The evidence for these PAMPs is, however, scarce.

Although the specific immune mechanisms are not clear, it is believed that microbial exposure may affect T lymphocytes which have an important function in controlling immune responses, including help for B cell production of antibodies (IgE, IgG, IgA, IgM). T-helper-2 (Th_2) cells stimulate B cells to produce IgE upon allergen stimulation, whereas T-helper-1 cells (Th_1) inhibit this process. The initial interpretation was that growing up in a more hygienic environment with less microbial exposure may enhance atopic (Th_2) immune responses, whereas microbial pressure would drive the response of the immune system – which is known to be skewed in an atopic Th_2 direction during fetal and perinatal life – into a Th_1 direction and away from its tendency to develop atopic immune responses. More recently, an alternative interpretation has been offered which involves inadequate immunoregulation by T regulatory cells wich control both Th_1 and Th_2 immune responses. Active regulation through T regulatory cells is believed to be critical in maintaining tolerance to allergens through a balanced Th_1/Th_2 immune response. The hypothesis is that a lack of microbial exposure may result in reduced immune suppression of T regulatory cells allowing upregulation of both Th_1 and Th_2 immunity, thus rendering subjects more susceptible to developing allergies (as well as Th_1

conditions including autoimmune disease which has also increased in prevalence in the past few decades). However, the immunological mechanisms underlying the observed epidemiological associations remain largely unclear (Romagnani 2004).

56.6.3.2 Animal Contact

Several studies have shown that the presence of pets in the home early in life is inversely associated with atopy in children (Hesselmar et al. 1999). Other studies have also shown a protective effect of pet ownership and asthma, for example, de Meer et al. (2004b) showed that having had a cat before the age of 18 protected against atopy to outdoor allergens, airway hyperreactivity, current wheeze, and current asthma. These results should, however, be interpreted with caution, since avoidance behavior (removal of pets in the families with sensitized and/or symptomatic children) may have contributed to this inverse association. However, in a longitudinal study in which subjects with childhood asthma at enrolment were excluded from the analyses, the protective effects actually increased (de Meer et al. 2004b), whereas a decrease would be expected if selective avoidance was a major issue. There are also studies that have found no association, or a positive association, between pet exposure and asthma, despite showing an inverse association with atopy (Apelberg et al. 2001; Kerkhof et al. 2009). In other parts of the world (Guinea-Bissau and Nepal), it has been shown that pigs and cattle in the home are associated with less atopy (Shaheen et al. 1996; Melsom et al. 2001). Thus, although the evidence for a protective effect on atopy is reasonably consistent, currently, it is unclear whether pets can also protect against asthma.

At present it is not clear which specific exposures and immunological mechanisms underlie the observed protective effects on atopy of animal contact, but increased microbial exposure may play a role, which would be consistent with the hygiene hypothesis (see above).

56.6.3.3 Farming

A number of studies have found consistently low prevalences of allergies and asthma in farmers' children, both in high-income countries such as Canada, the US, Australia, New Zealand, and Europe (Braun-Fahrlander et al. 1999; Ernst and Cormier 2000; Riedler et al. 2000, 2001; von Ehrenstein et al. 2000; Downs et al. 2001; Klintberg et al. 2001; Horak et al. 2002; Remes et al. 2002, 2003; Wickens et al. 2002; Chrischilles et al. 2004; Alfven et al. 2006; Dimich-Ward et al. 2006; Perkin and Strachan 2006; Midodzi et al. 2007; Douwes et al. 2008), and in low-income countries including Mongolia and Southern Africa (Weinberg 2000; Viinanen et al. 2007). These protective effects for allergies and asthma have also been observed in adult farmers (Leynaert et al. 2001; Kauffmann et al. 2002; Kilpelainen et al. 2002; Portengen et al. 2002, 2005; Braback et al. 2004; Eduard et al. 2004a, b; Koskela et al. 2005; Radon et al. 2006; Chen et al. 2007; Douwes et al. 2007; Smit et al. 2007, 2008) despite the increased risks of other respiratory conditions such as COPD, reduced lung function, and farmers' lung (Schenker et al. 1998). The fact that similar effects are found in low-income countries

(Weinberg 2000; Viinanen et al. 2007), where people have less opportunity to move away from farming because of allergies and asthma, further suggests that selection effects are unlikely to explain the often substantial lower risk in farming communities (Douwes et al. 2009).

The observed protective effects of farming on allergies and asthma have been particularly strong for animal contact (Riedler et al. 2000; von Ehrenstein et al. 2000; Downs et al. 2001; Remes et al. 2002, 2003; Douwes et al. 2008). Farm animals are associated with high exposures to microorganisms, in particular bacterial endotoxin (Douwes et al. 2002b). Also, an upregulation of several innate immune receptors specific for microbial products (TLRs and CD14) has been shown in farmers' children (Ege et al. 2006) suggesting that microorganisms and microbial products may be involved. In fact, a recent study showed that exposure to a wide variety of environmental microorganisms as well as exposures to specific fungal and bacterial species explained a substantial fraction of the inverse association between farm upbringing and asthma (Ege et al. 2011). Exposures early in life including the prenatal period appear particularly protective, although continued exposure may be required to maintain optimal protection (Douwes et al. 2007).

Consumption of unpasteurized farm milk in farmers' and non-farmers' children has also been shown to be protective in several of the farmers studies (Riedler et al. 2000; Barnes et al. 2001; Wickens et al. 2002; Perkin and Strachan 2006; Waser et al. 2007). The etiological mechanisms are unclear, but probiotic bacteria or other currently unidentified non-microbial components in farm milk may play a role. Evidence from other populations with anthroposophic lifestyles which are characterized by (among other things) diets rich in (probiotic) microbes (Alm et al. 1999) suggests that this protective effect may not be limited only to the farming environment, although the findings of these studies have not always been consistent (Alfven et al. 2006). These observations are also consistent with the hygiene hypothesis.

56.6.4 Can Protective Factors Explain the International Patterns and Time Trends?

As noted above, the hygiene hypothesis, if it is correct, would explain an increase in atopic/allergic asthma. However, since about one-half of asthma cases involve non-atopic/allergic mechanisms (Pearce et al. 1999), it is questionable whether the hygiene hypothesis on its own can explain the large increases observed over the last decades or the global prevalence patterns, particularly since there is some evidence that non-atopic asthma may have increased more than atopic asthma (Thomsen et al. 2004). Also, although housing conditions are unlikely to have become more hygienic in United States inner city populations, asthma prevalence has increased significantly in those populations, particularly among African Americans living in poverty (Crater et al. 2001). Finally, the hygiene hypothesis is unlikely to explain why asthma prevalence is now apparently falling in affluent countries, as exposures to factors that have previously been identified as being "protective" (family size, endotoxin exposure, infectious diseases, pets) are likely to have decreased in more recent times rather than increased. These findings thus further emphasize the potential limitations of the current hygiene hypothesis (Douwes and Pearce 2008). Nevertheless, whatever mechanism is involved, it is becoming increasingly clear that the "package" of changes associated with westernization may be contributing to the global increases in asthma susceptibility and prevalence.

56.7 The Role of Epidemiology in Respiratory Allergy and Asthma Research

In this final section, we will make some more general observations on the role of epidemiology in respiratory allergy and asthma research. Until fairly recently and with some notable exceptions, most asthma epidemiology was done by clinicians rather than professional epidemiologists, and epidemiology was largely regarded as a means of "confirming" clinically based observations and theories (e.g., that allergens can cause asthma exacerbations and therefore "must" also cause asthma itself) (Pearce and Douwes 2009). The field was perhaps comparable to the situation in cardiovascular disease and cancer in the 1950s. For these diseases, the first epidemiological step in exploring the etiology of these conditions involved descriptive epidemiology ("person, place, and time"), particularly international prevalence or incidence comparisons such as the MONICA project (Tuomilehto et al. 1987) and Cancer Incidence in Five Continents (Doll et al. 1966). These naturally led to the development of hypotheses as to the possible explanations for the observed global patterns, which were investigated in analytical studies (cohort and case-control studies), eventually evolving into more complex studies involving the use of biomarkers and genetic testing.

Because asthma epidemiology has arrived on the scene relatively late, there has been a natural tendency to jump straight into the use of relatively high-tech methods (e.g., BHR testing, atopy testing, IgE, airway eosinophil measurements, and more recently, genetic testing) focusing on very specific etiological mechanisms (e.g., TH_1/TH_2, eosinophilia), without the "reality check" which large-scale population comparisons provide. As a result, we have developed theories of asthma etiology (e.g., the TH_1/TH_2 theory, and the associated "hygiene hypothesis") which work well in mice, but do not seem to work so well in Latin America or inner city populations in the US (see Sect. 56.6.4) (Pearce and Douwes 2006). In the last couple of decades, epidemiology has played the major role in rectifying this situation, for example, by mounting standardized international prevalence surveys to provide global asthma maps, and other population comparisons (see Sect. 56.5), as a starting point for the more careful and appropriate development of theories of asthma etiology which not only work in the laboratory but which have the potential to explain what is happening in the population context. Of course, such population comparisons are not the solution by themselves, but they are an essential first step in the process of identifying and understanding the causes and etiology of asthma.

Much of the confusion that has permeated asthma epidemiology has simply reflected confusion about what asthma is and how it should be defined. Ironically, most clinicians have little problem diagnosing and treating asthma, which is regarded simply as “variable airways obstruction” (e.g., “asthma is a disease characterized by wide variation over short periods of time in resistance to flow in the airways of the lung” (American Thoracic Society Committee on Diagnostic Standards 1962)) – a condition that is easily diagnosed on the basis of a clinical symptom history, perhaps supplemented by serial measurements of lung function (Pearce et al. 1998a). Despite some major therapeutic disasters (Pearce et al. 1998b), symptoms are relatively easy to control (at least for the eosinophilic phenotype; see Sect. 56.3.1). However, there has not been the same clarity in understanding the etiological mechanisms by which variable airways obstruction occurs, or in the methods used for measuring asthma in prevalence studies (Pekkanen and Pearce 1999). Epidemiology can make an important contribution in the development of a better understanding of the etiology of asthma and whether it constitutes several different conditions or a single condition. This in turn will help to redefine the best methods of defining asthma in clinical practice and in epidemiological surveys.

56.7.1 The Future of Allergy and Asthma Epidemiology

Although epidemiological research may not have yielded a target for asthma intervention (yet), it has made a major contribution in changing the focus from risk factors to protective factors and identifying populations with a “naturally” low prevalence of allergies and asthma. These observations provide interesting new avenues of research with a high likelihood of identifying novel targets for prevention and treatment.

So where could innovative new theories of the etiology of asthma come from? Global prevalence (and if possible, incidence) comparisons can play a major role in this process (Pearce 1999) as has been successfully demonstrated in asthma studies such as ISAAC (Asher et al. 1995) and ECRHS (Burney et al. 1994), but, clearly, they are just part of a larger scientific process, and these sorts of initiatives are now well established for many non-communicable diseases such as cancer (Doll et al. 1966) and coronary heart disease (WHO MONICA Project Principal Investigators 1988).

Perhaps the most promising new direction is the development of truly interdisciplinary research that integrates the usually separate worlds of epidemiology, social science and biomedical, and clinical research. Traditionally, these disciplines have been conducting research in relative isolation, a situation which has distinct disadvantages. In particular, (i) many animal models of disease are only partially applicable to human populations; (ii) clinical studies are generally not well equipped to determine the causal exposures of primary causation; and (iii) epidemiological studies often do not fully acknowledge the complexity of the biological responses involved.

Mice do not get asthma, and even after manipulation to produce an "asthma model," they do not exhibit the pathophysiology that is consistent with asthma in humans (Wenzel and Holgate 2006). Thus, although we have learnt a great deal about immunology in mice in the past few decades, these mouse models have provided only limited understanding of the underlying immunological mechanisms of human asthma. Nonetheless, most biomedical research on asthma continues to be conducted in mice. Similarly, the overemphasis of allergen exposure as the "causal" factor of asthma, as well as the lack of appreciation of the heterogeneity of asthma in many epidemiological studies, has guided biomedical and clinical research into studying (almost exclusively) allergic mechanisms and allergy-specific treatment options, with very little consideration of other potentially relevant mechanisms. In particular, there are now several studies suggesting that corticosteroid treatment is less or non-effective in asthma phenotypes that do not involve allergic mechanisms and subsequent non-eosinophilic airway inflammation (Berry et al. 2007), despite these phenotypes being very common (Douwes et al. 2002a; Simpson et al. 2007). Also, studies of asthma genetics continue to be based on populations of mixed phenotypes limiting the potential to find meaningful results. This "tunnel vision" has led to assumptions that findings in Western countries can be extrapolated to the rest of the world, and it has taken international collaborations such as the ISAAC study, to show that, for example, the strong associations between atopy and asthma symptoms that have been repeatedly observed in Western countries are not so evident in low- and middle-income countries (Weinmayr et al. 2007). Thus, considering the complex interplay between environmental exposures, genetic susceptibility, immunological mechanisms, as well as social and cultural factors involved in asthma development – and indeed most non-communicable diseases – an interdisciplinary approach bringing together expertise from each of these disciplines is most desirable, as it is likely to open up new avenues of cutting edge research, yielding greater explanatory power, more efficient use of research funding, and more efficient translation into disease prevention. Also, a closer link with biomedical research will ensure the development of more sensitive and valid methods that can be widely applied in population studies to improve both disease definition (including the assessment and recognition of disease phenotypes) and relevant exposures.

Epidemiologists have a major role to play in coordinating, integrating, and expanding these efforts, and that the "population perspective" (Pearce 1996, 1999) provides a valid reference point and "reality check" for such interdisciplinary work. However, although epidemiology is well placed to take the lead in developing interdisciplinary research, we also see a role for more "traditional" epidemiology. In particular, epidemiology has a strong tradition of generating new hypotheses or refuting old dogmas. Studies such as ISAAC (see Sect. 56.5.3) are an excellent example where the use of a one-page questionnaire conducted in more than one million children worldwide has made a major impact on how we think about asthma etiology (Enarson 2005).

56.8 Conclusions

Despite decades of intensive biomedical and epidemiological research, the etiology and pathogenesis of asthma is still poorly understood. In particular, a large number of potential risk factors for asthma have been identified including genetic factors, allergen exposure, demographic parameters, diet, obesity, indoor and outdoor pollution, passive and active tobacco smoking, occupational exposures, viral infections, stress/anxiety, and the use of paracetamol (acetaminophen). However, none of these risk factors on their own appears to explain the substantial global increases in asthma prevalence observed in the last few decades. They also cannot explain the significant differences in asthma prevalence between countries. Recent studies have shown that the increase in asthma prevalence appears to have leveled off in many high-income countries, with some even showing a decrease. The reasons for this are also unclear. As a consequence, a single target for intervention has not (yet) been identified.

Understanding why these changes in prevalence are occurring, and ascertaining which elements of the "package" of twentieth century economic development and lifestyle changes are responsible, is essential in order to develop effective intervention programs to halt the current global asthma epidemic. Recent epidemiological studies have provided innovative theories of the etiology of asthma, such as the hygiene hypothesis, which have considerable potential to guide the development of feasible primary (and secondary) prevention options. However, given the complex gene-environment interactions (Vercelli 2009) and diverse asthma phenotypes (Douwes et al. 2002a), any single type of intervention is unlikely to prevent all or even a substantial proportion of asthma cases – a "one size fits all" approach is unlikely to work. Epidemiology (alongside other disciplines) has a major role to play in developing new theories regarding asthma causation and the subsequent development and evaluation of effective prevention strategies. However, for these efforts to be maximally effective, they would need to be part of a multidisciplinary program that includes expertise (in addition to epidemiology) in biomedical, clinical, genetic, psychological, and other disciplines.

Acknowledgements The Centre for Public Health Research is supported by a Programme Grant from the Health Research Council (HRC) of New Zealand.

References

Aberg N (1989) Asthma and allergic rhinitis in Swedish conscripts. Clin Exp Allergy 19:59–63

Adams R, Ruffin R, Wakefield M, Campbell D, Smith B (1997) Asthma prevalence, morbidity and management practices in South Australia, 1992–1995. Aust N Z J Med 27:672–679

Aggarwal AN, Gupta D, Jindal SK (2006) The relationship between FEV1 and peak expiratory flow in patients with airways obstruction is poor. Chest 130:1454–1461

Akerman MJ, Calacanis CM, Madsen MK (2004) Relationship between asthma severity and obesity. J Asthma 41:521–526

Alfven T, Braun-Fahrlander C, Brunekreef B, von Mutius E, Riedler J, Scheynius A, van Hage M, Wickman M, Benz MR, Budde J, Michels KB, Schram D, Ublagger E, Waser M, Pershagen G (2006) Allergic diseases and atopic sensitization in children related to farming and anthroposophic lifestyle – the PARSIFAL study. Allergy 61:414–421

Alm JS, Lilja G, Pershagen G, Scheynius A (1998) BCG vaccination does not seem to prevent atopy in children with atopic heredity. Allergy 53:537

Alm JS, Swartz J, Lilja G, Scheynius A, Pershagen G (1999) Atopy in children of families with an anthroposophic lifestyle. Lancet 353:1485–1488

Almqvist C, Garden F, Xuan W, Mihrshahi S, Leeder SR, Oddy W, Webb K, Marks GB (2007) Omega-3 and omega-6 fatty acid exposure from early life does not affect atopy and asthma at age 5 years. J Allergy Clin Immunol 119:1438–1444

Alving K, Weitzberg E, Lundberg JM (1993) Increased amount of nitric oxide in exhaled air of asthmatics. Eur Respir J 6:1368–1370

American Lung Association (2010) Trends in asthma morbidity and mortality. http://www.lungusa.org. Accessed 27 Aug 2011

American Thoracic Society (1987) Standardization of spirometry – 1987 update. Statement of the American Thoracic Society. Am Rev Respir Dis 136:1285–1298

American Thoracic Society Committee on Diagnostic Standards (1962) Definitions and classification of chronic bronchitis, asthma and pulmonary emphysema. Am Rev Resp Dis 85:762–768

Anderson HR, Pottier AC, Strachan DP (1992) Asthma from birth to age 23 – incidence and relation to prior and concurrent atopic disease. Thorax 47:537–542

Anderson HR, Butland BK, Strachan DP (1994) Trends in prevalence and severity of childhood asthma. BMJ 308:1600–1604

Anderson HR, Ruggles R, Strachan DB, Austin JB, Burr M, Jeffs D, Standring P, Steriu A, Goulding R (2004) Trends in prevalence of symptoms of asthma, hay fever, and eczema in 12–14 year olds in the British Isles, 1995–2002: questionnaire survey. BMJ 328:1052–1053

Anto JM, Sunyer J, Newman Taylor AJ (1996) Comparison of soybean epidemic asthma and occupational asthma. Thorax 51:743–749

Apelberg BJ, Aoki Y, Jaakkola JJ (2001) Systematic review: exposure to pets and risk of asthma and asthma-like symptoms. J Allergy Clin Immunol 107:455–460

Armstrong BK, White E, Saracci R (1992) Principles of exposure measurements in epidemiology. Oxford University Press, New York

Asher MI, Keil U, Anderson HR, Beasley R, Crane J, Martinez F, Mitchell EA, Pearce N, Sibbald B, Stewart AW (1995) International Study of Asthma and Allergies in Childhood (ISAAC): rationale and methods. Eur Respir J 8:483–491

Asher MI, Anderson HR, Stewart AW, Crane J, Ait-Khaled N, Anabwani G, Anderson HR, Beasley R, Björkstén B, Burr ML, Clayton TO, Crane J, Ellwood P, Keil U, Lai CKW, Mallol J, Martinez FD, Mitchell EA, Montefort S, Pearce N, Robertson CF, Shah JR, Sibbald B, Strachan DP, Weiland SK, Williams HC (1998) Worldwide variations in the prevalence of asthma symptoms: International Study of Asthma and Allergies in Childhood (ISAAC). Eur Respir J 12:315–335

Asher MI, Montefort S, Bjorksten B, Lai CKW, Strachan DP, Weiland SK, Williams H (2006) Worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and eczema in childhood: ISAAC Phases One and Three repeat multicountry cross-sectional surveys. Lancet 368:733–743

Auerbach I, Springer C, Godfrey S (1993) Total population survey of the frequency and severity of asthma in 17 year old boys in an urban area in Israel. Thorax 48:139–141

Ball TM, Castro-Rodriguez JA, Griffith KA, Holberg CJ, Martinez FD, Wright AL (2000) Siblings, day-care attendance, and the risk of asthma and wheezing during childhood. N Engl J Med 343:538–543

Barnes M, Cullinan P, Athanasaki P, MacNeill S, Hole AM, Harris J, Kalogeraki S, Chatzinikolaou M, Drakonakis N, Bibaki-Liakou V, Taylor AJN, Bibakis I (2001) Crete: does farming explain urban-rural differences in atopy? Clin Exp Allergy 31:1822–1828

Barr RG, Wentowski CC, Curhan GC, Somers SC, Stampfer MJ, Schwartz J, Speizer FE, Camargo CA Jr (2004) Prospective study of acetaminophen use and newly diagnosed asthma among women. Am J Respir Crit Care Med 169:836–841

Beasley R, Keil U, von Mutius E, Pearce N, Ait-Khaled N, Anabwani G, Anderson HR, Asher MI, Björkstén B, Burr ML, Clayton TO, Crane J, Ellwood P, Lai CKW, Mallo LJ, Martinez FD, Mitchell EA, Montefort S, Robertson CF, Shah JR, Sibbald B, Stewart AW, Strachan DP, Weiland SK, Williams HC (1998) Worldwide variation in prevalence of symptoms of asthma, allergic rhinoconjunctivitis and atopic eczema: ISAAC. Lancet 351:1225–1232

Beasley R, Clayton T, Crane J, von Mutius E, Lai CK, Montefort S, Stewart A (2008) Association between paracetamol use in infancy and childhood, and risk of asthma, rhinoconjunctivitis, and eczema in children aged 6–7 years: analysis from Phase Three of the ISAAC programme. Lancet 372:1039–1048

Berry MA, Shaw DE, Green RH, Brightling C, Wardlaw AJ, Pavord ID (2005) The use of exhaled nitric oxide concentration to identify eosinophilic airway inflammation: an observational study in adults with asthma. Clin Exp Allergy 35:1175–1179

Berry MA, Morgan A, Shaw DE, Parker D, Green R, Brightling C, Bradding P, Wardlaw AJ, Pavord ID (2007) Pathological features and inhaled corticosteroid response of eosinophilic and non-eosinophilic asthma. Thorax 62:1043–1049

Bjerg A, Sandström T, Lundbäck B, Rönmark E (2010) Time trends in asthma and wheeze in Swedish children 1996–2006: prevalence and risk factors by sex. Allergy 65:48–55

Blackley C (1873) Experimental research in the causes and nature of catarrhus aestivus (hay fever or hay asthma). Baillière Tindall and Cox, London

Boezen HM, van der Zee SC, Postma DS, Vonk JM, Gerritsen J, Hoek G, Brunekreef B, Rijcken B, Schouten JP (1999) Effects of ambient air pollution on upper and lower respiratory symptoms and peak expiratory flow in children. Lancet 353:874–878

Bollag U, Capkun G, Caesar J, Low N (2005) Trends in primary care consultations for asthma in Switzerland, 1989–2002. Int J Epidemiol 34:1012–1018

Bolte G, Bischof W, Borte M, Lehmann I, Wichmann HE, Heinrich J (2003) Early endotoxin exposure and atopy development in infants: results of a birth cohort study. Clin Exp Allergy 33:770–776

Bottcher MF, Bjorksten B, Gustafson S, Voor T, Jenmalm MC (2003) Endotoxin levels in Estonian and Swedish house dust and atopy in infancy. Clin Exp Allergy 33:295–300

Braback L, Hjern A, Rasmussen F (2004) Trends in asthma, allergic rhinitis and eczema among Swedish conscripts from farming and non-farming environments. A nationwide study over three decades. Clin Exp Allergy 34:38–43

Braback L, Hjern A, Rasmussen F (2005) Body mass index, asthma and allergic rhinoconjunctivitis in Swedish conscripts - a national cohort study over three decades. Respir Med 99:1010–1014

Brauer M, Hoek G, Smit HA, de Jongste JC, Gerritsen J, Postma DS, Kerkhof M, Brunekreef B (2007) Air pollution and development of asthma, allergy and infections in a birth cohort. Eur Respir J 29:879–888

Braun-Fahrlander C, Gassner M, Grize L, Neu U, Sennhauser FH, Varonier HS, Vuille JC, Wuthrich B (1999) Prevalence of hay fever and allergic sensitization in farmer's children and their peers living in the same rural community. Clin Exp Allergy 29:28–34

Braun-Fahrlander C, Riedler J, Herz U, Eder W, Waser M, Grize L, Maisch S, Carr D, Gerlach F, Bufe A, Lauener RP, Schierl R, Renz H, Nowak D, von Mutius E (2002) Environmental exposure to endotoxin and its relation to asthma in school-age children. N Engl J Med 347: 869–877

Brenner JS, Kelly CS, Wenger AD, Brich SM, Morrow AL (2001) Asthma and obesity in adolescents: is there an association? J Asthma 38:509–515

Buchele G, Genuneit J, Weinmayr G, Bjorksten B, Gehring U, von Mutius E, Priftanji A, Stein RT, Addo-Yobo EO, Priftis KN, Shah JR, Forastiere F, Svabe V, Crane J, Nystad W, Garcia-Marcos L, Saraclar Y, El-Sharif N, Strachan DP (2010) International variations in bronchial responsiveness in children: findings from ISAAC phase two. Pediatr Pulmonol 45:796–806

Burney P (2002) The changing prevalence of asthma? Thorax 57(Suppl 2):II36–II39

Burney PG, Britton JR, Chinn S, Tattersfield AE, Papacosta AO, Kelson MC, Anderson F, Corfield DR (1987) Descriptive epidemiology of bronchial reactivity in an adult population: results from a community study. Thorax 42:38–44

Burney PG, Laitinen LA, Perdrizet S, Huckauf H, Tattersfield AE, Chinn S, Poisson N, Heeren A, Britton JR, Jones T (1989) Validity and repeatability of the IUATLD (1984) Bronchial Symptoms Questionnaire: an international comparison. Eur Respir J 2:940–945

Burney P, Chinn S, Rona RJ (1990) Has the prevalence of asthma increased in children? Evidence from a national study of health and growth 1973–86. BMJ 300:1306–1310

Burney PG, Luczynska C, Chinn S, Jarvis D (1994) The European Community Respiratory Health Survey. Eur Respir J 7:954–960

Burney P, Chinn S, Luczynska C, Jarvis D, Vermeire P, Bousquet J, Nowak D, Prichard J, deMarco R, Rijcken B, Anto J, Alves J, Boman G, Kesteloot H, Nielsen N, Paoletti P (1996) Variations in the prevalence of respiratory symptoms, self-reported asthma attacks, and use of asthma medication in the European Community Respiratory Health Survey (ECRHS). Eur Respir J 9:687–695

Burr ML, Butland BK, King S, Vaughan-Williams E (1989) Changes in asthma prevalence: two surveys 15 years apart. Arch Dis Child 64:1452–1456

Burr ML, Limb ES, Maguire MJ, Amarah L, Eldridge BA, Layzell JCM, Merrett TG (1993) Infant-feeding, wheezing, and allergy – a prospective-study. Arch Dis Childhood 68:724–728

Bustos P, Amigo H, Oyarzun M, Rona RJ (2005) Is there a causal relation between obesity and asthma? Evidence from Chile. Int J Obes 29:804–809

Butland BK, Strachan DP, Anderson HR (1997) The home environment and asthma symptoms in childhood: two population based case-control studies 13 years apart. Thorax 52:618–624

Campbell D, Ruffin R, Mcevoy R, Crockett A (1992) South Australian asthma prevalence survey. Aust NZ Med J 22:A658 (abstract)

Celedon JC, Litonjua AA, Ryan L, Weiss ST, Gold DR (2002) Lack of association between antibiotic use in the first year of life and asthma, allergic rhinitis, or eczema at age 5 years. Am J Respir Crit Care Med 166:72–75

Celedon JC, Fuhlbrigge A, Rifas-Shiman S, Weiss ST, Finkelstein JA (2004) Antibiotic use in the first year of life and asthma in early childhood. Clin Exp Allergy 34:1011–1016

Chan-Yeung MJ, Malo L (1994) Epidemiology of occupational asthma. In: Busse W, Holgate ST (eds) Asthma and rhinitis. Blackwell, Oxford, pp 44–57

Checkoway H, Pearce N, Kriebel D (2004) Research methods in occupational epidemiology. Oxford University Press, New York

Chen Y, Rennie D, Cormier Y, McDuffie H, Pahwa P, Dosman J (2007) Reduced risk of atopic sensitization among farmers: The Humboldt Study. Int Arch Allergy Immun 144:338–342

Chinn S, Jarvis D, Burney P, Luczynska C, Ackermann-Liebrich U, Anto JM, Cerveri I, De Marco R, Gislason T, Heinrich J, Janson C, Kunzli N, Leynaert B, Neukirch F, Schouten J, Sunyer J, Svanes C, Vermeire P, Wjst M (2004) Increase in diagnosed asthma but not in symptoms in the European Community Respiratory Health Survey. Thorax 59:646–651

Chrischilles E, Ahrens R, Kuehl A, Kelly K, Thorne P, Burmeister L, Merchant J (2004) Asthma prevalence and morbidity among rural Iowa schoolchildren. J Allergy Clin Immunol 113:66–71

Ciba Foundation Guest Symposium (1959) Terminology definitions, classification of chronic pulmonary emphysema and related conditions. Thorax 14:286–299

Ciprandi G, Vizzaccaro A, Cirillo I, Crimi P, Canonica GW (1996) Increase of asthma and allergic rhinitis prevalence in young Italian men. Int Arch Allergy Immunol 111:278–283

Cohet C, Cheng S, MacDonald C, Baker M, Foliaki S, Huntington N, Douwes J, Pearce N (2004) Infections, medication use, and the prevalence of symptoms of asthma, rhinitis, and eczema in childhood. J Epidemiol Community Health 58:852–857

Cookson H, Granell R, Joinson C, Ben-Shlomo Y, Henderson AJ (2009) Mothers' anxiety during pregnancy is associated with asthma in their children. J Allergy Clin Immunol 123:847–853

Corver K, Kerkhof M, Brussee JE, Brunekreef B, van Strien RT, Vos AP, Smit HA, Gerritsen J, Neijens HJ, de Jongste JC (2006) House dust mite allergen reduction and allergy at 4 yr: follow up of the PIAMA-study. Pediatr Allergy Immunol 17:329–336

Crater DD, Heise S, Perzanowski M, Herbert R, Morse CG, Hulsey TC, Platts-Mills T (2001) Asthma hospitalization trends in Charleston, South Carolina, 1956 to 1997: twenty-fold increase among black children during a 30-year period. Pediatrics 108:E97

Czebe K, Barta I, Antus B, Valyon M, Horvath I, Kullmann T (2008) Influence of condensing equipment and temperature on exhaled breath condensate pH, total protein and leukotriene concentrations. Respir Med 102:720–725

de Meer G, Marks GB, Postma DS (2004a) Direct or indirect stimuli for bronchial challenge testing: what is the relevance for asthma epidemiology? Clin Exp Allergy 34:9–16

de Meer G, Toelle BG, Ng K, Tovey E, Marks GB (2004b) Presence and timing of cat ownership by age 18 and the effect on atopy and asthma at age 28. J Allergy Clin Immunol 113:433–438

de Meer G, Marks GB, de Jongste JC, Brunekreef B (2005) Airway responsiveness to hypertonic saline: dose-response slope or PD15? Eur Respir J 25:153–158

Department of Health and Senior Services, State of New Jersey (2006) Industries and asthma-causing agents. http://www.state.nj.us/health/eoh/survweb/wra/agents.shtml. Accessed 27 Aug 2011

Department of Human Health and Human Services (DHHS), US (2006) The health consequences of involuntary exposure to tobacco smoke: a report of the Surgeon General. US Department of Human Health and Human Services, Centres for Disease Control and Prevention, Coordinating Centre for Health Promotion, Office on Smoking and Health, Atlanta

Devereux G (2007) Early life events in asthma – diet. Pediatr Pulmonol 42:663–673

Dimich-Ward H, Chow Y, Chung J, Trask C (2006) Contact with livestock – a protective effect against allergies and asthma? Clin Exp Allergy 36:1122–1129

Dodge RR, Burrows B (1980) The prevalence and incidence of asthma and asthma-like symptoms in a general population sample. Am Rev Respir Dis 122:567–575

Doll R, Payne P, Waterhouse J (eds) (1966) Cancer incidence in five continents. Springer, Berlin

Dolovich J, Hargreave F (1981) The asthma syndrome: inciters, inducers, and host characteristics. Thorax 36:614–644

Douwes J, Pearce N (2002) Asthma and the westernization ‘package’. Int J Epidemiol 31: 1098–1102

Douwes J, Pearce N (2003) Is indoor mold exposure a risk factor for asthma? Am J Epidemiol 158:203–206

Douwes J, Pearce N (2008) Commentary: the end of the hygiene hypothesis? Int J Epidemiol 37:570–572

Douwes J, Gibson P, Pekkanen J, Pearce N (2002a) Non-eosinophilic asthma: importance and possible mechanisms. Thorax 57:643–648

Douwes J, Pearce N, Heederik D (2002b) Does environmental endotoxin exposure prevent asthma? Thorax 57:86–90

Douwes J, Le Gros G, Gibson P, Pearce N (2004) Can bacterial endotoxin exposure reverse atopy and atopic disease? J Allergy Clin Immunol 114:1051–1054

Douwes J, van Strien R, Doekes G, Smit J, Kerkhof M, Gerritsen J, Postma D, de Jongste J, Travier N, Brunekreef B (2006) Does early indoor microbial exposure reduce the risk of asthma? The Prevention and Incidence of Asthma and Mite Allergy birth cohort study. J Allergy Clin Immunol 117:1067–1073

Douwes J, Travier N, Huang K, Cheng S, McKenzie J, Le Gros G, von Mutius E, Pearce N (2007) Lifelong farm exposure may strongly reduce the risk of asthma in adults. Allergy 62:1158–1165

Douwes J, Cheng S, Travier N, Cohet C, Niesink A, McKenzie J, Cunningham C, Le Gros G, von Mutius E, Pearce N (2008) Farm exposure in utero may protect against asthma, hay fever and eczema. Eur Respir J 32:603–611

Douwes J, Brooks C, Pearce N (2009) The protective effects of farming on allergies and asthma: have we learnt anything since 1873? Exp Rev Clin Immunol 5:213–219

Douwes J, Brooks C, Pearce N (2010) Stress and asthma: hippocrates revisited. J Epidemiol Community Health 64:561–562

Douwes J, Brooks C, Pearce N (2011) Asthma nervosa: old concept, new insight. Eur Respir J 37:986–990

Downs SH, Marks GB, Mitakakis TZ, Leuppi JD, Car NG, Peat JK (2001) Having lived on a farm and protection against allergic diseases in Australia. Clin Exp Allergy 31:570–575

Dowse GK, Turner KJ, Stewart GA, Alpers MP, Woolcock AJ (1985) The association between Dermatophagoides mites and the increasing prevalence of asthma in village communities within the Papua New Guinea highlands. J Allergy Clin Immunol 75:75–83

Dressel H, de la Motte D, Reichert J, Ochmann U, Petru R, Angerer P, Holz O, Nowak D, Jorres RA (2008) Exhaled nitric oxide: independent effects of atopy, smoking, respiratory tract infection, gender and height. Respir Med 102:962–969

Dring T, Harper C, Leigh J (1689) Practice of physick, pharmaceutica rationalis or the operations of medicine in humane bodies. London

Eagan TM, Bakke PS, Eide GE, Gulsvik A (2002) Incidence of asthma and respiratory symptoms by sex, age and smoking in a community study. Eur Respir J 19:599–605

Edfors-Lubs ML (1971) Allergy in 7000 twin pairs. Acta Allergologica 26:249–285

Eduard W, Douwes J, Omenaas E, Heederik D (2004a) Do farming exposures cause or prevent asthma? Results from a study of adult Norwegian farmers. Thorax 59:381–386

Eduard W, Omenaas E, Bakke PS, Douwes J, Heederik D (2004b) Atopic and non-atopic asthma in a farming and a general population. Am J Ind Med 46:396–399

Ege MJ, Bieli C, Frei R, van Strien RT, Riedler J, Ublagger E, Schram-Bijkerk D, Brunekreef B, van Hage M, Scheynius A, Pershagen G, Benz MR, Lauener R, von Mutius E, Braun-Fahrlander C (2006) Prenatal farm exposure is related to the expression of receptors of the innate immunity and to atopic sensitization in school-age children. J Allergy Clin Immunol 117:817–823

Ege MJ, Mayer M, Normand AC, Genuneit J, Cookson WO, Braun-Fahrländer C, Heederik D, Piarroux R, von Mutius E; GABRIELA Transregio 22 Study Group (2011) Exposure to environmental microorganisms and childhood asthma. N Engl J Med 364:701–709

Ehrlich RI, Du Toit D, Jordaan E, Volmink JA, Weinberg EG, Zwarenstein M (1995) Prevalence and reliability of asthma symptoms in primary school children in Cape Town. Int J Epidemiol 24:1138–1145

Ellison-Loschmann L, Pearce N (2006) Improving access to health care among New Zealand's Maori population. Am J Pub Health 96:612–617

Ellison-Loschmann L, Sunyer J, Plana E, Pearce N, Zock JP, Jarvis D, Janson C, Anto JM, Kogevinas M (2007) Socioeconomic status, asthma and chronic bronchitis in a large community-based study. Eur Respir J 29:897–905

Ellul-Micallef R (1976) Asthma: a look at the past. Br J Dis Chest 70:112–116

Ellwood P, Asher MI, Bjorksten M, Burr M, Pearce N, Robertson CF (2001) Diet and asthma, allergic rhinoconjunctivitis and atopic eczema symptom prevalence: an ecological analysis of the International Study of Asthma and Allergies in Childhood (ISAAC) data. ISAAC Phase One Study Group. Eur Respir J 17:436–443

Ellwood P, Asher MI, Beasley R, Clayton TO, Stewart AW (2005) The International Study of Asthma and Allergies in Childhood (ISAAC): phase three rationale and methods. Int J Tuberc Lung Dis 9:10–16

Enarson D (2005) Fostering a spirit of critical thinking: the ISAAC story. Int J Tuberc Lung Dis 9:1

Environmental Protection Agency (EPA), US (1992) Respiratory health effects of passive smoking: lung cancer and other disorders. Office of Research and Development, US Environmental Protection Agency, Washington, DC

Ernst P, Cormier Y (2000) Relative scarcity of asthma and atopy among rural adolescents raised on a farm. Am J Respir Crit Care Med 161:1563–1566

Farber HJ, Wattigney W, Berenson G (1997) Trends in asthma prevalence: the Bogalusa Heart Study. Ann Allergy Asthma Immunol 78:265–269

Farooqi IS, Hopkin JM (1998) Early childhood infection and atopic disorder. Thorax 53:927–932

Fishwick D, Pearce N, D'Souza W, Lewis S, Town I, Armstrong R, Kogevinas M, Crane J (1997) Occupational asthma in New Zealanders: a population based study [comment]. Occup Environ Med 54:301–306

Fleming DM, Crombie DL (1987) Prevalence of asthma and hay fever in England and Wales. BMJ 294:279–283

Flohr C, Weiland SK, Weinmayr G, Bjorksten B, Braback L, Brunekreef B, Buchele G, Clausen M, Cookson WO, von Mutius E, Strachan DP, Williams HC (2008) The role of atopic sensitization in flexural eczema: findings from the International Study of Asthma and Allergies in Childhood phase two. J Allergy Clin Immunol 121:141–147

Foliaki S, Pearce N, Bjorksten B, Mallol J, Montefort S, von Mutius E (2009) Antibiotic use in infancy and symptoms of asthma, rhinoconjunctivitis, and eczema in children 6 and 7 years old: International Study of Asthma and Allergies in Childhood phase III. J Allergy Clin Immunol 124:982–989

Friedman NJ, Zeiger RS (2005) The role of breast-feeding in the development of allergies and asthma. J Allergy Clin Immunol 115:1238–1248

Garcia-Marcos L, Quiros AB, Hernandez GG, Guillen-Grima F, Diaz CG, Urena IU, Pena AA, Monge RB, Suarez-Varela MM, Varela AL, Cabanillas PG, Garrido JB (2004) Stabilization of asthma prevalence among adolescents and increase among schoolchildren (ISAAC phases I and III) in Spain. Allergy 59:1301–1307

Gehring U, Bolte G, Borte M, Bischof W, Fahlbusch B, Wichmann HE, Heinrich J (2001) Exposure to endotoxin decreases the risk of atopic eczema in infancy: a cohort study. J Allergy Clin Immunol 108:847–854

Gergen PJ, Mullally DI, Evans R (1988) National survey of prevalence of asthma among children in the United States, 1976 to 1980. Pediatrics 81:1–7

Gibson PG, Henry RL, Thomas P (2000) Noninvasive assessment of airway inflammation in children: induced sputum, exhaled nitric oxide, and breath condensate. Eur Respir J 16: 1008–1015

Gilliland FD, Berhane K, Islam T, McConnell R, Gauderman WJ, Gilliland SS, Avol E, Peters JM (2003) Obesity and the risk of newly diagnosed asthma in school-age children. Am J Epidemiol 158:406–415

Global Initiative for Asthma (GINA) (2006) Global strategy for asthma management and prevention. Global initiative for asthma. http://www.ginasthma.org. Accessed 27 Aug 2011

Global Initiative for Asthma (GINA) (2010) Global strategy for asthma management and prevention. Global initiative for asthma. http://www.ginasthma.org. Accessed 5 Oct 2011

Godfrey KM, Barker DJP, Osmond C (1994) Disproportionate fetal growth and raised IgE concentration in adult life. Clin Exp Allergy 24:641–648

Gold DR, Damokosh AI, Dockery DW, Berkey CS (2003) Body-mass index as a predictor of incident asthma in a prospective cohort of children. Pediatr Pulmonol 36:514–521

Gotzsche PC, Hammarquist C, Burr M (1998) House dust mite control measures in the management of asthma: meta-analysis. BMJ 317:1105–1110

Haahtela T, Lindholm H, Bjorksten F, Koskenvuo K, Laitinen LA (1990) Prevalence of asthma in Finnish young men. BMJ 301:266–268

Hakala K, Stenius-Aarniala B, Sovijarvi A (2000) Effects of weight loss on peak flow variability, airways obstruction, and lung volumes in obese patients with asthma. Chest 118:1315–1321

Hesselmar B, Aberg N, Aberg B, Eriksson B, Bjorksten B (1999) Does early exposure to cat or dog protect against later allergy development? Clin Exp Allergy 29:611–617

Hill R, Williams J, Tattersfield A, Britton J (1989) Change in use of asthma as a diagnostic label for wheezing illness in schoolchildren. BMJ 299:898

Hirsch T, Range U, Walther KU, Hederer B, Lassig S, Frey G, Leupold W (1998) Prevalence and determinants of house dust mite allergen in East German homes. Clin Exp Allergy 28:956–964

Hoffmeyer F, Raulf-Heimsoth M, Bruning T (2009) Exhaled breath condensate and airway inflammation. Curr Opin Allergy Clin Immunol 9:16–22

Holgate ST (2004) The epidemic of asthma and allergy. J R Soc Med 97:103–110

Horak F, Studnicka M, Gartner C, Veiter A, Tauber E, Urbanek R, Frischer T (2002) Parental farming protects children against atopy: longitudinal evidence involving skin prick tests. Clin Exp Allergy 32:1155–1159

Hsieh K-H, Shen J-J (1991) Prevalence of childhood asthma in Taipei, Taiwan and other Asian Pacific countries. J Asthma 25:73–82

Infante-Rivard C, Esnaola Sukia S, Roberge D, Baumgarten M (1987) The changing frequency of childhood asthma. J Asthma 24:283–288

James A, Ryan G (1997) Testing airway responsiveness using inhaled methacholine or histamine. Respirology 2:97–105

Jatakanon A, Lim S, Kharitonov SA, Chung KF, Barnes PJ (1998) Correlation between exhaled nitric oxide, sputum eosinophils, and methacholine responsiveness in patients with mild asthma. Thorax 53:91–95

Jenkins HS, Devalia JL, Mister RL, Bevan AM, Rusznak C, Davies RJ (1999) The effect of exposure to ozone and nitrogen dioxide on the airway response of atopic asthmatics to inhaled allergen: dose- and time-dependent effects. Am J Respir Crit Care Med 160:33–39

Johnston SL (2007) Innate immunity in the pathogenesis of virus-induced asthma exacerbations. Proc Am Thorac Soc 4:267–270

Johnston SL, Pattemore PK, Sanderson G, Smith S, Lampe F, Josephs L, Symington P, Otoole S, Myint SH, Tyrrell DAJ, Holgate ST (1995) Community study of role of viral-infections in exacerbations of asthma in 9–11 year-old children. BMJ 310:1225–1229

Josephs LK, Gregg I, Holgate ST (1990) Does non-specific bronchial responsiveness indicate the severity of asthma? Eur Respir J 3:220–227

Karjalainen A, Kurppa K, Martikainen R, Klaukka T, Karjalainen J (2001) Work is related to a substantial portion of adult-onset asthma incidence in the Finnish population. Am J Respir Crit Care Med 164:565–568

Kauffmann F, Oryszczyn MP, Maccario J (2002) The protective role of country living on skin prick tests, immunoglobulin E and asthma in adults from the epidemiological study on the genetics and environment of asthma, bronchial hyper-responsiveness and atopy. Clin Exp Allergy 32:379–386

Keeney EL (1964) The history of asthma from Hippocrates to Meltzer. J Allergy 35:215–226

Kerkhof M, Wijga AH, Brunekreef B, Smit HA, de Jongste JC, Aalberse RC, Hoekstra MO, Gerritsen J, Postma DS (2009) Effects of pets on asthma development up to 8 years of age: the PIAMA study. Allergy 64:1202–1208

Kharitonov SA, Barnes PJ (2006) Exhaled biomarkers. Chest 130:1541–1546

Kharitonov SA, Robbins RA, Yates D, Keatings V, Barnes PJ (1995) Acute and chronic effects of cigarette smoking on exhaled nitric oxide. Am J Respir Crit Care Med 152:609–612

Kharitonov SA, Yates DH, Barnes PJ (1996) Inhaled glucocorticoids decrease nitric oxide in exhaled air of asthmatic patients. Am J Respir Crit Care Med 153:454–457

Kilpelainen M, Terho EO, Helenius H, Koskenvuo M (2002) Childhood farm environment and asthma and sensitization in young adulthood. Allergy 57:1130–1135

Kimbell-Dunn M, Pearce N, Beasley R (1999) Asthma. In: Goldman M, Hatch M (eds) Women and health. Academic, San Diego, pp 724–739

Klintberg B, Berglund N, Lilja G, Wickman M, van Hage-Hamsten M (2001) Fewer allergic respiratory disorders among farmers' children in a closed birth cohort from Sweden. Eur Respir J 17:1151–1157

Kogevinas M, Anto JM, Soriano JB, Tobias A, Burney P, Martinez Moratalla J, Almar E, Aguilar X, Arevalo M, Mateos A, Sanchez A, Teixido A, Vizcaya M, Sunyer J, Burgos F, Castellsague J, Galobardes MB, Benavides FG, Roca J, Muniozguren N, Errezola M, Capelastegui A, Ramos J, Maldonado JA, Sanchez JL, Pereira A, Gravalos J, Quiros R, Azofra J, Palenciano L, Payo F, Rego G, Vega A (1996) The risk of asthma attributable to occupational exposures – a population-based study in Spain. Am J Respir Crit Care Med 154:137–143

Koskela HO, Happonen KK, Remes ST, Pekkanen J (2005) Effect of farming environment on sensitisation to allergens continues after childhood. Occup Environ Med 62:607–611

Kramer MS, Matush L, Vanilovich I, Platt R, Bogdanovich N, Sevkovskaya Z, Dzikovich I, Shishko G, Mazer B (2007) Effect of prolonged and exclusive breast feeding on risk of allergy and asthma: cluster randomised trial. BMJ 335:815

Kummeling I, Stelma FF, Dagnelie PC, Snijders BE, Penders J, Huber M, van Ree R, van den Brandt P, Thijs C (2007) Early life exposure to antibiotics and the subsequent development of eczema, wheeze, and allergic sensitization in the first 2 years of life: the KOALA Birth Cohort Study. Pediatrics 119:e225–e231

Lau S, Illi S, Sommerfeld C, Niggemann B, Bergmann R, von Mutius E, Wahn U (2000) Early exposure to house-dust mite and cat allergens and development of childhood asthma: a cohort study. Multicentre Allergy Study Group. Lancet 356:1392–1397

Lau S, Nickel R, Niggemann B, Gruber C, Sommerfeld C, Illi S, Kulig M, Forster J, Wahn U, Groeger M, Zepp F, Kamin W, Bieber I, Tacke U, Wahn V, Bauer CP, Bergmann R, von Mutius E (2002) The development of childhood asthma: lessons from the German Multicentre Allergy Study (MAS). Paediatr Respir Rev 3:265–272

Lemanske RF Jr, Jackson DJ, Gangnon RE, Evans MD, Li Z, Shult PA, Kirk CJ, Reisdorf E, Roberg KA, Anderson EL, Carlson-Dakes KT, Adler KJ, Gilbertson-White S, Pappas TE, Dasilva DF, Tisler CJ, Gern JE (2005) Rhinovirus illnesses during infancy predict subsequent childhood wheezing. J Allergy Clin Immunol 116:571–577

Leung R, Ho P, Lam CWK, Lai CWK (1997) Sensitization to inhaled allergens as a risk factor for asthma and allergic diseases in Chinese population. J Allergy Clin Immunol 99: 594–599

Leynaert B, Neukirch C, Jarvis D, Chinn S, Burney P, Neukirch F (2001) Does living on a farm during childhood protect against asthma, allergic rhinitis, and atopy in adulthood? Am J Respir Crit Care Med 164:1829–1834

Liard R, Chansin R, Neukirch F, Levallois M, Leproux P (1988) Prevalence of asthma among teenagers attending school in Tahiti. J Epidemiol Community Health 42:149–151

Littlejohns L, Macdonald LD (1993) The relationship between severe asthma and social-class. Respir Med 87:139–143

Littlejohns P, Ebrahim S, Anderson R (1989) Prevalence and diagnosis of chronic respiratory symptoms in adults. BMJ 298:1556–1560

Lowe A, Abramson M, Dharmage S, Allen K (2009) Paracetamol as a risk factor for allergic disorders. Lancet 373:120; author reply: 120–121

Magnus PJ, Jaakkola JK (1997) Secular trend in the occurrence of asthma among children and young adults: critical appraisal of repeated cross sectional surveys. BMJ 314:1795–1799

Manfreda J, Becker AB, Wang PZ, Roos LL, Anthonisen NR (1993) Trends in physician-diagnosed asthma prevalence in Manitoba between 1980 and 1990. Chest 103:151–157

Mannino DM, Mott J, Ferdinands JM, Camargo CA, Friedman M, Greves HM, Redd SC (2006) Boys with high body masses have an increased risk of developing asthma: findings from the National Longitudinal Survey of Youth (NLSY). Int J Obes 30:6–13

Martinez FD (1997) Complexities of the genetics of asthma. Am J Respir Crit Care Med 156: S117–S122

Martinez FD, Morgan WJ, Wright AL, Holberg CJ, Taussig LM (1988) Diminished lung-function as a predisposing factor for wheezing respiratory illness in infants. N Engl J Med 319: 1112–1117

Matheson MC, Erbas B, Balasuriya A, Jenkins MA, Wharton CL, Tang ML, Abramson MJ, Walters EH, Hopper JL, Dharmage SC (2007) Breast-feeding and atopic disease: a cohort study from childhood to middle age. J Allergy Clin Immunol 120:1051–1057

Matricardi PM, Rosmini F, Ferrigno L, Nisini R, Rapicetta M, Chionne P, Stroffolini T, Pasquini P, D'Amelio R (1997) Cross sectional retrospective study of prevalence of atopy among Italian military students with antibodies against hepatitis A virus. BMJ 314:999–1003

Matricardi PM, Rosmini F, Riondino S, Fortini M, Ferrigno L, Rapicetta M, Bonini S (2000) Exposure to foodborne and orofecal microbes versus airborne viruses in relation to atopy and allergic asthma: epidemiological study. BMJ 320:412–417

McConnell R, Berhane K, Gilliland F, London SJ, Islam T, Gauderman WJ, Avol E, Margolis HG, Peters JM (2002) Asthma in exercising children exposed to ozone: a cohort study. Lancet 359:386–391

McDonnell WF, Abbey DE, Nishino N, Lebowitz MD (1999) Long-term ambient ozone concentration and the incidence of asthma in nonsmoking adults: the AHSMOG Study. Environ Res 80:110–121

McKeever TM, Lewis SA, Smith C, Collins J, Heatlie H, Frischer M, Hubbard R (2002a) Early exposure to infections and antibiotics and the incidence of allergic disease: a birth cohort study with the West Midlands General Practice Research Database. J Allergy Clin Immunol 109: 43–50

McKeever TM, Lewis SA, Smith C, Hubbard R (2002b) The importance of prenatal exposures on the development of allergic disease: a birth cohort study using the West Midlands General Practice Database. Am J Respir Crit Care Med 166:827–832

Melsom T, Brinch L, Hessen JO, Schei MA, Kolstrup N, Jacobsen BK, Svanes C, Pandey MR (2001) Asthma and indoor environment in Nepal. Thorax 56:477–481

Mendall MA, Kumar D (1998) Antibiotic use, childhood affluence and irritable bowel syndrome (IBS). Eur J Gastroenterol Hepatol 10:59–62

Meredith S, Nordman H (1996) Occupational asthma: measures of frequency from four countries. Thorax 51:435–440

Meredith SK, Taylor VM, McDonald JC (1991) Occupational respiratory disease in the United Kingdom 1989: a report to the British Thoracic Society and the Society of Occupational Medicine by the SWORD project group. Br J Ind Med 48:292–298

Midodzi WK, Rowe BH, Majaesic CM, Senthilselvan A (2007) Reduced risk of physician-diagnosed asthma among children dwelling in a farming environment. Respirology 12:692–699

Miller BD, Wood BL, Lim J, Ballow M, Hsu C (2009) Depressed children with asthma evidence increased airway resistance: "vagal bias" as a mechanism? J Allergy Clin Immunol 124:66–73

Michel S, Liang L, Depner M, Klopp N, Ruether A, Kumar A, Schedel M, Vogelberg C, von Mutius E, von Berg A, Bufe A, Rietschel E, Heinzmann A, Laub O, Simma B, Frischer T, Genuneit J, Gut IG, Schreiber S, Lathrop M, Illig T, Kabesch M (2010) Unifying candidate gene and GWAS approaches in asthma. PLoS One 5:e13894

Mitchell EA (1983) Increasing prevalence of asthma in children. N Z Med J 96:463–464

Mitchell EA, Stewartt AW, IPOS Group (2002) The ecological relationship of tobacco smoking to the prevalence of symptoms of asthma and other atopic diseases in children: the International Study of Asthma and Allergies in Childhood (ISAAC). Eur J Epidemiol 17:667–673

Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, von Mutius E, Farrall M, Lathrop M, Cookson WO, GABRIEL Consortium (2010) A large-scale, consortium-based genomewide association study of asthma. N Engl J Med 363:1211–1221

Mommers M, Guekjens-Sijstermans C, Swaen GMH, van Schayck CP (2005) Trends in the prevalence of respiratory symptoms and treatment in Dutch children over a 12 year period: results of the fourth consecutive survey. Thorax 60:97–99

Moncayo AL, Vaca M, Oviedo G, Erazo S, Quinzo I, Fiaccone RL, Chico ME, Barreto ML, Cooper PJ (2010) Risk factors for atopic and non-atopic asthma in a rural area of Ecuador. Thorax 65:409–416

Morrison Smith J (1976) The prevalence of asthma and wheezing in children. Br J Dis Chest 70:73–77

Newson RB, Shaheen SO, Chinn S, Burney PG (2000) Paracetamol sales and atopic disease in children and adults: an ecological analysis. Eur Respir J 16:817–823

Ninan TK, Russell G (1992) Respiratory symptoms and atopy in Aberdeen schoolchildren: evidence from two surveys 25 years apart. BMJ 304:873–875

Nishima S (1993) A study on the prevalence of bronchial asthma in school children in western districts of Japan – comparison between the studies in 1982 and in 1992 with the same methods and same districts. The Study Group of the Prevalence of Bronchial Asthma, the West Japan Study Group of Bronchial Asthma. Arerugi 42:192–204

Ober C, Yao TC (2011) The genetics of asthma and allergic disease: a 21st century perspective. Immunol Rev 242:10–30

Office of Environmental Health Hazard Assessment (OEHHA) (1997) Health effects of exposure to environmental tobacco smoke California. Office of Environmental Health Hazard Assessment, California Environmental Protection Agency

Olin AC, Rosengren A, Thelle DS, Lissner L, Bake B, Toren K (2006) Height, age, and atopy are associated with fraction of exhaled nitric oxide in a large adult general population sample. Chest 130:1319–1325

Omran M, Russell G (1996) Continuing increase in respiratory symptoms and atopy in Aberdeen schoolchildren. BMJ 312:34

Osler W (1892) Bronchial asthma. Appleton, New York

Pattemore PK, Ellison-Loschmann L, Asher MI, Barry DMJ, Clayton TO, Crane J, D'Souza WJ, Ellwood P, Ford RPK, Mackay RJ, Mitchell EA, Moyes C, Pearce N, Stewart AW (2004) Asthma prevalence in European, Maori, and Pacific children in New Zealand: ISAAC study. Pediatr Pulmonol 37:433–442

Pearce N (1992) Methodological problems of time-related variables in occupational cohort studies. Rev Epidemiol Sante Publique 40(Suppl) 1:S43–S54

Pearce N (1996) Traditional epidemiology, modern epidemiology, and public health. Am J Public Health 86:678–683

Pearce N (1999) Epidemiology as a population science. Int J Epidemiol 28:S1015–S1018

Pearce N, Beasley R (1999) Measuring morbidity in adult asthmatics. Int J Tuberc Lung Dis 3:185–191

Pearce N, Douwes J (2005) Asthma time trends – mission accomplished? Int J Epidemiol 34: 1018–1019

Pearce N, Douwes J (2006) The global epidemiology of asthma in children. Int J Tuberc Lung Dis 10:125–132

Pearce N, Douwes J (2009) Response: time for species–course epidemiology? Int J Epidemiol 38:403–410

Pearce N, Weiland S, Keil U, Langridge P, Anderson HR, Strachan D, Bauman A, Young L, Gluyas P, Ruffinet D (1993) Self-reported prevalence of asthma symptoms in children in Australia, England, Germany and New Zealand: an international comparison using the ISAAC protocol. Eur Respir J 6:1455–1461

Pearce N, de Sanjose S, Boffetta P, Kogevinas M, Saracci R, Savitz D (1995) Limitations of biomarkers of exposure in cancer epidemiology. Epidemiology 6:190–194

Pearce N, Beasley R, Burgess C, Crane J (1998a) Asthma epidemiology: principles and methods. Oxford University Press, New York

Pearce N, Beasley R, Crane J, Burgess C (1998b) Pharmacoepidemiology of asthma deaths. In: Tilson H (ed) Pharmacoepidemiology: an introduction, 3rd edn. Harvey Whitney, Cincinatti, pp 473–494

Pearce N, Pekkanen J, Beasley R (1999) How much asthma is really attributable to atopy? Thorax 54:268–272

Pearce N, Beasley R, Pekkanen J (2000a) Role of bronchial responsiveness testing in asthma prevalence surveys. Thorax 55:352–354

Pearce N, Douwes J, Beasley R (2000b) The rise and rise of asthma: a new paradigm for the new millennium? J Epidemiol Biostat 5:5–16

Pearce N, Douwes J, Beasley R (2000c) Is allergen exposure the major primary cause of asthma? Thorax 55:424–431

Pearce N, Ait-Khaled N, Beasley R, Mallol J, Keil U, Mitchell E, Robertson C (2007) Worldwide trends in the prevalence of asthma symptoms: phase III of the International Study of Asthma and Allergies in Childhood (ISAAC). Thorax 62:758–766

Peat JK, Tovey E, Toelle BG, Haby MM, Gray EJ, Mahmic A, Woolcock AJ (1996) House dust mite allergens. A major risk factor for childhood asthma in Australia. Am J Respir Crit Care Med 153:141–146

Peckham C, Butler N (1978) A national study of asthma in childhood. J Epidemiol Community Health 32:79–85

Pekkanen J, Pearce N (1999) Defining asthma in epidemiological studies. Eur Respir J 14:951–957

Perdrizet S, Neukirch F, Cooreman J, Liard R (1987) Prevalence of asthma in adolescents in various parts of France and its relationship to respiratory allergic manifestations. Chest 91:104S–106S

Perkin MR, Strachan DP (2006) Which aspects of the farming lifestyle explain the inverse association with childhood allergy? J Allergy Clin Immunol 117:1374–1381

Pijnenburg MW, Bakker EM, Hop WC, De Jongste JC (2005) Titrating steroids on exhaled nitric oxide in children with asthma: a randomized controlled trial. Am J Respir Crit Care Med 172:831–836

Porsbjerg C, Lund TK, Pedersen L, Backer V (2009) Inflammatory subtypes in asthma are related to airway hyperresponsiveness to mannitol and exhaled NO. J Asthma 46:606–612

Portengen L, Sigsgaard T, Omland O, Hjort C, Heederik D, Doekes G (2002) Low prevalence of atopy in young Danish farmers and farming students born and raised on a farm. Clin Exp Allergy 32:247–253

Portengen L, Preller L, Tielen M, Doekes G, Heederik D (2005) Endotoxin exposure and atopic sensitization in adult pig farmers. J Allergy Clin Immunol 115:797–802

Priftanji A, Strachan D, Burr M, Sinamati J, Shkurti A, Grabocka E, Kaur B, Fitzpatrick S (2001) Asthma and allergy in Albania and the UK. Lancet 358:1426–1427

Quanjer PH, Tammeling GJ, Cotes JE, Pedersen OF, Peslin R, Yernault JC (1993) Lung volumes and forced ventilatory flows. Report Working Party Standardization of Lung Function Tests, European Community for Steel and Coal. Official statement of the European respiratory society. Eur Respir J Suppl 16:5–40

Radon K, Schulze A, Nowak D (2006) Inverse association between farm animal contact and respiratory allergies in adulthood: protection, underreporting or selection? Allergy 61:443–446

Redline S, Gold D (1994) Challenges in interpreting gender differences in asthma. Am J Respir Crit Care Med 150:1219–1221

Reisman J, Schachter HM, Dales RE, Tran K, Kourad K, Barnes D, Sampson M, Morrison A, Gaboury I, Blackman J (2006) Treating asthma with omega-3 fatty acids: where is the evidence? A systematic review. BMC Complement Altern Med 6:26

Remes ST, Pekkanen J, Soininen L, Kajosaari M, Husman T, Koivikko A (2002) Does heredity modify the association between farming and allergy in children? Acta Paediatr 91:1163–1169

Remes ST, Iivanainen K, Koskela H, Pekkanen J (2003) Which factors explain the lower prevalence of atopy amongst farmers' children? Clin Exp Allergy 33:427–434

Riedler J, Eder W, Oberfeld G, Schreuer M (2000) Austrian children living on a farm have less hay fever, asthma and allergic sensitization. Clin Exp Allergy 30:194–200

Riedler J, Braun-Fahrlander C, Eder W, Schreuer M, Waser M, Maisch S, Carr D, Schierl R, Nowak D, von Mutius E (2001) Exposure to farming in early life and development of asthma and allergy: a cross-sectional survey. Lancet 358:1129–1133

Robertson CF, Heycock E, Bishop J, Nolan T, Olinsky A, Phelan PD (1991) Prevalence of asthma in Melbourne schoolchildren: changes over 26 years. BMJ 302:1116–1118

Robertson CF, Roberts MF, Kappers JH (2004) Asthma prevalence in Melbourne schoolchildren: have we reached the peak? Med J Aust 180:273–276

Romagnani S (2004) Immunologic influences on allergy and the TH1/TH2 balance. J Allergy Clin Immunol 113:395–400

Ronchetti R, Rennerova Z, Barreto M, Villa MP (2007) The prevalence of atopy in asthmatic children correlates strictly with the prevalence of atopy among nonasthmatic children. Int Arch Allergy Immunol 142:79–85

Salter HH (1860) On asthma: its pathology and treatment. John Churchill & Sons, London

Sandford A, Weir T, Pare P (1996) The genetics of asthma. Am J Respir Crit Care Med 153: 1749–1765

Schenker MB, Christiani D, Cormier Y, Dimich-Ward H, Doekes G, Dosman J, Douwes J, Dowling K, Enarson D, Green F, Heederik D, Husman K, Kennedy S, Kullman G, Lacasse Y, Lawson B, Malmberg P, May J, McCurdy S, Merchant J, Myers J, Nieuwenhuijsen M, Olenchock S, Saiki C, Schwartz D, Seiber J, Thorne P, Wagner G, White N, Xu XP, Chan-Yeung M (1998) Respiratory health hazards in agriculture. Am J Resp Crit Care Med 158:S1–S76

Sears MR, Lewis S, Herbison GP, Robson B, Flannery EM, Holdaway MD, Pearce N, Crane J, Silva PA (1997) Comparison of reported prevalences of recent asthma in longitudinal and cross-sectional studies. Eur Respir J 10:51–54

Sears MR, Greene JM, Willan AR, Taylor DR, Flannery EM, Cowan JO, Herbison JP, Poulton R (2002) Long-term relation between breastfeeding and development of atopy and asthma in children and young adults: a longitudinal study. Lancet 360:901–907

Shaheen SO, Aaby P, Hall AJ, Barker DJ, Heyes CB, Shiell AW, Goudiaby A (1996) Measles and atopy in Guinea-Bissau. Lancet 347:1792–1796

Shaheen SO, Sterne JA, Songhurst CE, Burney PG (2000) Frequent paracetamol use and asthma in adults. Thorax 55:266–270

Shaheen SO, Newson RB, Sherriff A, Henderson AJ, Heron JE, Burney PG, Golding J (2002) Paracetamol use in pregnancy and wheezing in early childhood. Thorax 57:958–963

Shaheen SO, Newson RB, Henderson AJ, Headley JE, Stratton FD, Jones RW, Strachan DP (2005) Prenatal paracetamol exposure and risk of asthma and elevated immunoglobulin E in childhood. Clin Exp Allergy 35:18–25

Shaheen S, Potts J, Gnatiuc L, Makowska J, Kowalski ML, Joos G, van Zele T, van Durme Y, De Rudder I, Wohrl S, Godnic-Cvar J, Skadhauge L, Thomsen G, Zuberbier T, Bergmann KC, Heinzerling L, Gjomarkaj M, Bruno A, Pace E, Bonini S, Fokkens W, Weersink EJ, Loureiro C, Todo-Bom A, Villanueva CM, Sanjuas C, Zock JP, Janson C, Burney P (2008) The relation between paracetamol use and asthma: a GA2LEN European case-control study. Eur Respir J 32:1231–1236

Shaheen SO, Newson RB, Smith GD, Henderson AJ (2010) Prenatal paracetamol exposure and asthma: further evidence against confounding. Int J Epidemiol 39:790–794

Shaw RA, Crane J, O'Donnell TV, Porteous LE, Coleman ED (1990) Increasing asthma prevalence in a rural New Zealand adolescent population: 1975–89. Arch Dis Child 65:1319–1323

Shaw R, Woodman K, Crane J, Moyes C, Kennedy J, Pearce N (1994) Risk factors for asthma symptoms in Kawerau children. N Z Med J 107:387–391

Shaw R, Woodman K, Ayson M, Dibdin S, Winkelmann R, Crane J, Beasley R, Pearce N (1995) Measuring the prevalence of bronchial hyper-responsiveness in children. Int J Epidemiol 24:597–602

Shore SA (2007) Obesity and asthma: lessons from animal models. J Appl Physiol 102:516–528

Sigurs N, Gustafsson PM, Bjarnason R, Lundberg F, Schmidt S, Sigurbergsson F, Kjellman B (2005) Severe respiratory syncytial virus bronchiolitis in infancy and asthma and allergy at age 13. Am J Respir Crit Care Med 171:137–141

Simpson JL, Scott R, Boyle MJ, Gibson PG (2006) Inflammatory subtypes in asthma: assessment and identification using induced sputum. Respirology 11:54–61

Simpson JL, Grissell TV, Douwes J, Scott RJ, Boyle MJ, Gibson PG (2007) Innate immune activation in neutrophilic asthma and bronchiectasis. Thorax 62:211–218

Siroux V, Pin I, Oryszczyn MP, Le Moual N, Kauffmann F (2000) Relationships of active smoking to asthma and asthma severity in the EGEA study. Epidemiological study on the genetics and environment of asthma. Eur Respir J 15:470–477

Smit LAM, Zuurbier M, Doekes G, Wouters IM, Heederik D, Douwes J (2007) Hay fever and asthma symptoms in conventional and organic farmers in The Netherlands. Occup Environ Med 64:101–107

Smit LAM, Heederik D, Doekes G, Blom C, van Zweden I, Wouters IM (2008) Exposure-response analysis of allergy and respiratory symptoms in endotoxin-exposed adults. Eur Respir J 31:1241–1248

Souza da Cunha S, Barreto ML, Fiaccone RL, Cooper PJ, Alcantara-Neves NM, Simões Sde M, Cruz AA, Rodrigues LC (2010) Asthma cases in childhood attributed to atopy in tropical area in Brazil. Rev Panam Salud Publica 28:405–411

Sporik R, Holgate ST, Plattsmills TAE, Cogswell JJ (1990) Exposure to house-dust mite Allergen (Der-P-I) and the development of asthma in childhood – a prospective study. N Engl J Med 323:502–507

Stein RT, Sherrill D, Morgan WJ, Holberg CJ, Halonen M, Taussig LM, Wright AL, Martinez FD (1999) Respiratory syncytial virus in early life and risk of wheeze and allergy by age 13 years. Lancet 354:541–545

Stewart AW, Mitchell EA, Pearce N, Strachan DP, Weilandon SK, ISAAC Steering Committee (2001) The relationship of per capita gross national product to the prevalence of symptoms of asthma and other atopic diseases in children (ISAAC). Int J Epidemiol 30:173–179

Strachan DP (1989) Hay fever, hygiene, and household size. BMJ 299:1259–1260

Strachan DP, Cook DG (1998) Parental smoking and childhood asthma: longitudinal and case-control studies. Thorax 53:204–212

Sunyer J, Jarvis D, Pekkanen J, Chinn S, Janson C, Leynaert B, Luczynska C, Garcia-Esteban R, Burney P, Anto JM (2004) Geographic variations in the effect of atopy on asthma in the European Community Respiratory Health Study. J Allergy Clin Immunol 114:1033–1039

Taylor DR, Mandhane P, Greene JM, Hancox RJ, Filsell S, McLachlan CR, Williamson AJ, Cowan JO, Smith AD, Sears MR (2007) Factors affecting exhaled nitric oxide measurements: the effect of sex. Respir Res 8:82

Tepas EC, Litonjua AA, Celedon JC, Sredl D, Gold DR (2006) Sensitization to aeroallergens and airway hyperresponsiveness at 7 years of age. Chest 129:1500–1508

Thomsen SF, Ulrik CS, Larsen K, Backer V (2004) Change in prevalence of asthma in Danish children and adolescents. Ann Allergy Asthma Immunol 92:506–511

Toelle BG, Peat JK, Salome CM, Mellis CM, Woolcock AJ (1992) Toward a definition of asthma for epidemiology. Am Rev Resp Disease 146:633–637

Travers J, Marsh S, Aldington S, Williams M, Shirtcliffe P, Pritchard A, Weatherall M, Beasley R (2007) Reference ranges for exhaled nitric oxide derived from a random community survey of adults. Am J Respir Crit Care Med 176:238–242

Tuomilehto J, Kuulasmaa K, Torppa J (1987) WHO MONICA Project: geographic variation in mortality from cardiovascular diseases. World Health Stat Q 40:171–184

Turner MO, Hussack P, Sears MR, Dolovich J, Hargreave FE (1995) Exacerbations of asthma without sputum eosinophilia. Thorax 50:1057–1061

Unger L, Harris MC (1974) Stepping stones in allergy. Ann Allergy 32:214–230

van Strien RT, Verhoeff AP, Brunekreef B, Van Wijnen JH (1994) Mite antigen in house dust: relationship with different housing characteristics in The Netherlands. Clin Exp Allergy 24:843–853

Varner AE, Busse WW, Lemanske RE Jr (1998) Hypothesis: decreased use of pediatric aspirin has contributed to the increasing prevalence of childhood asthma. Ann Allergy Asthma Immunol 81:347–351

Vercelli D (2009) Gene-environment interactions: the road less traveled by in asthma genetics. J Allergy Clin Immunol 123:26–27

Vig RS, Forsythe P, Vliagoftis H (2006) The role of stress in asthma: insight from studies on the effect of acute and chronic stressors in models of airway inflammation. Ann N Y Acad Sci 1088:65–77

Viinanen A, Munhbayarlah S, Zevgee T, Narantsetseg L, Naidansuren T, Koskenvuo M, Helenius H, Terho EO (2007) The protective effect of rural living against atopy in Mongolia. Allergy 62:272–280

von Ehrenstein OS, von Mutius E, Illi S, Baumann L, Bohm O, von Kries R (2000) Reduced risk of hay fever and asthma among children of farmers. Clin Exp Allergy 30:187–193

von Hertzen LC (2002) Maternal stress and T-cell differentiation of the developing immune system: possible implications for the development of asthma and atopy. J Allergy Clin Immunol 109:923–928

von Mutius E, Martinez FD, Fritzsch C (1994) Skin test reactivity and number of siblings. BMJ 308:692–695

von Mutius E, Weiland SK, Fritzsch C, Duhme H, Keil U (1998) Increasing prevalence of hay fever and atopy among children in Leipzig, East Germany. Lancet 351:862–866

von Mutius E, Schwartz J, Neas LM, Dockery D, Weiss ST (2001) Relation of body mass index to asthma and atopy in children: the National Health and Nutrition Examination Study III. Thorax 56:835–838

Vork KL, Broadwin RL, Blaisdell RJ (2007) Developing asthma in childhood from exposure to secondhand tobacco smoke: insights from a meta-regression. Environ Health Perspect 115:1394–1400

Waser M, Michels KB, Bieli C, Floistrup H, Pershagen G, von Mutius E, Ege M, Riedler J, Schram-Bijkerk D, Brunekreef B, van Hage M, Lauener R, Braun-Fahrlander C (2007) Inverse association of farm milk consumption with asthma and allergy in rural and suburban populations across Europe. Clin Exp Allergy 37:661–670

Watson JP, Cowen P, Lewis RA (1996) The relationship between asthma admission rates, routes of admission, and socioeconomic deprivation. Eur Respir J 9:2087–2093

Weeke E (1992) Epidemiology of allergic diseases in children. Rhinology 30(Suppl 13):5–12

Weiland SK, Pearce N (2004) Asthma prevalence in adults: good news? Thorax 59:637–638

Weinberg EG (2000) Urbanization and childhood asthma: an African perspective. J Allergy Clin Immunol 105:224–231

Weinmayr G, Weiland SK, Bjorksten B, Brunekreef B, Buchele G, Cookson WO, Garcia-Marcos L, Gotua M, Gratziou C, van Hage M, von Mutius E, Riikjarv MA, Rzehak P, Stein RT, Strachan DP, Tsanakas J, Wickens K, Wong GW (2007) Atopic sensitization and the international variation of asthma symptom prevalence in children. Am J Respir Crit Care Med 176: 565–574

Weitzman M, Gortmaker SL, Sobol AM, Perrin JM (1992) Recent trends in the prevalence and severity of childhood asthma. JAMA 268:2673–2677

Wenzel S, Holgate ST (2006) The mouse trap: it still yields few answers in asthma. Am J Respir Crit Care Med 174:1173–1176

Whincup PH, Cook DG, Strachan DP, Papacosta O (1993) Time trends in respiratory symptoms in childhood over a 24 year period. Arch Dis Child 68:729–734

World Health Organization (WHO) (2005) Air quality guidelines, global update 2005. Particulate matter, ozone, nitrogen dioxide and sulphur dioxide. WHO Regional Office for Europe, Copenhagen

World Health Organization (WHO) (2007) Global surveillance, prevention and control of chronic respiratory diseases: a comprehensive approach. World Health Organization, Geneva

World Health Organization (WHO) MONICA Project Principal Investigators (1988) The World Health Organization MONICA project (monitoring trends and determinants in cardiovascular disease): a major international collaboration. J Clin Epidemiol 41:105–114

Wickens K, Lane JM, Fitzharris P, Siebers R, Riley G, Douwes J, Smith T, Crane J (2002) Farm residence and exposures and the risk of allergic diseases in New Zealand children. Allergy 57:1171–1179

Wjst M, Hoelscher B, Frye C, Wichmann HE, Dold S, Heinrich J (2001) Early antibiotic treatment and later asthma. Eur J Med Res 6:263–271

Wright RJ, Fay ME, Suglia SF, Clark CJ, Evans JS, Dockery DW, Behbehani J (2010) War-related stressors are associated with asthma risk among older Kuwaitis following the 1990 Iraqi invasion and occupation. J Epidemiol Community Health 64:630–635

Yunginger JW, Reed CE, O’Connell EJ, Melton LJ 3rd, O’Fallon WM, Silverstein MD (1992) A community-based study of the epidemiology of asthma. Incidence rates, 1964–1983. Am Rev Respir Dis 146:888–894

Zamel N, McClean PA, Sandell PR, Siminovitch KA, Slutsky AS, Balter M, Canny G, Chapman K, Dzyngel B, Kesten S, Reisman J, Tarlo S, Urch B (1996) Asthma on Tristan de Cunha: looking for the genetic link. Am J Respir Crit Care Med 153:1902–1906

Epidemiology of Dental Diseases 57

David I. Conway, Alex D. McMahon, Douglas Robertson, and Lorna M. D. Macpherson

Contents

D.I. Conway (✉) • A.D. McMahon • D. Robertson • L.M.D. Macpherson
University of Glasgow Dental School, Glasgow, UK

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_51,

57.1 Introduction

Oral health means more than "good teeth" – it is integral to general health, is essential for well-being and is a determinant for quality of life.

Having good oral health allows us to speak, smile, kiss, touch, taste, chew, swallow, and even to cry or express ourselves. Conversely, oral diseases restrict activities and amount to significant time lost in school, at home, and in work. Further, it is the psychological impact of these diseases that can diminish the quality of life (Petersen 2003a). This wider definition of oral health should not detract from the importance of the two most important oral diseases – dental caries and periodontal disease, which will be the focus here, in addition to considering the general oral health measure of tooth loss. Other oral diseases and conditions including dental abscess, dental fluorosis, dental trauma, dental erosion, and dentofacial anomalies, in addition to oral mucosal conditions including oral cancer, are beyond the scope of this chapter.

Both dental caries and periodontal disease have the potential to be effectively prevented and treated, but have to be increasingly considered in the context of their strong relationship between socioeconomic factors. The incidence of oral disease is greatest in people from deprived areas and backgrounds (Sweeney et al. 1999; Selwitz et al. 2007; McMahon et al. 2010).

The integral interrelationship between oral health and general health is well documented, with oral disease and non-communicable chronic disease having many common risk factors (Sheiham and Watt 2000). Moreover, many general conditions have oral manifestations or can affect dental treatment.

Dental epidemiology is defined here to encapsulate epidemiology of oral diseases. Knowledge about the burden, trends, and causes of oral disease comes from the findings of dental epidemiology. Such understanding provides a basis for assessing and determining needs in populations, communities, and individuals. This, in turn, can support the planning and monitoring of the effectiveness of services and interventions designed to meet these oral health needs. Epidemiology provides the basis of preventive interventions and dental public health policies and interventions.

This chapter will discuss epidemiology in relation to describing the trends in dental caries, periodontal disease, and tooth loss and the analytical approaches to determining risk factors and the etiology of these diseases.

57.2 Definitions and Basic Concepts

57.2.1 Dental Caries

57.2.1.1 Pathogenesis

Dental caries or dental decay is a disease of the dental hard tissue. The pathogenesis of caries is a complex interaction between host factors (including teeth and saliva), oral bacteria, and fermentable carbohydrate in the form of sugars. Acid is produced as a by-product of the metabolism of dietary fermentable carbohydrate and oral bacteria within dental plaque. The resultant drop in pH at the tooth surface leads to the diffusion of calcium and phosphate ions from the enamel (demineralization); this process is reversed as the pH rises again (remineralization). Thus, caries is a dynamic disease process between periods of demineralization and remineralization, which progresses if demineralization predominates over time, and in turn will lead to loss of mineral content of enamel and eventually to carious cavitation. The disease develops in both the crowns and roots of teeth (Fejerskov 1997; Kidd 1987).

Enamel caries – is the early "white spot" lesion that appears on the surface of the tooth (the enamel). Within the enamel, this appearance is due to demineralization of the subsurface layer with the outer surface remaining more mineralized. With continued "acid attack" demineralization predominates and the outer surface changes from smooth to rough and can become stained, eventually becoming pitted or cavitated. In the early stages of demineralization, the process can be reversed through reuptake of calcium, phosphate, and fluoride ions (Selwitz et al. 2007).

Dentine (or dentinal) caries – is when caries reaches the dentine, which is the hard tissue surrounding the dental pulp and underneath the dental enamel. Caries is more severe by the time it has progressed to dentine. Dentinal caries is characterized by demineralization and subsequent bacterial invasion (Kidd and Fejerskov 2004).

Root-surface caries – appears when the root of the tooth becomes exposed due to recession of the gingival margins (the gums) from around the base of the crown of the tooth. This can result from poor oral hygiene, loss of periodontal attachment, and age; dental plaque can build up in these areas, and caries lesions can readily develop (Fejerskov and Kidd 2003).

57.2.1.2 Clinical Diagnosis

Dental caries is a continuous progress of increasing severity in terms of tooth destruction – from molecular changes to obvious frank cavitation. Therefore, the assessment of the presence of dental caries depends on the diagnostic criteria employed. In the clinical setting, dental practitioners would usually undertake a visual inspection of all visible surfaces of the teeth using a good light source, a dental mirror, and a dental probe – although the use of a dental probe is controversial and discouraged (James et al. 2001). Dental radiographs or fiber optic transillumination (FOTI) are also used to identify caries which may not be readily visible, particularly caries on the "interproximal" surfaces between teeth. White spot lesions can also be identified by blowing air across the tooth surface to remove saliva and make the "white spot lesions" more visible (Pitts 1997). Newer and more sensitive research methods and emerging technologies for caries diagnosis include laser fluorescence, light fluorescence, digital imaging fiber optic transillumination, and ultrasound that are potentially able to detect even more subclinical initial lesions, which are in a state of dynamic progression and regression at the early stage of the disease process, before they are discernible by conventional clinical methods. This gives the potential for lesions to be detected and for the impact of preventive care to be assessed in order to ensure that cavitation is avoided (Stookey and González-Cabezas 2001).

57.2.1.3 Epidemiological Measurement

In dental epidemiology, dental caries is usually diagnosed as caries into dentine (dentinal caries). These dentinal caries lesions are recorded following visual clinical inspection – examination with the naked eye under standard lighting, without the use of a dental probe or radiographs (Ismail 2004). Due to the possibility of subclinical (not visually obvious decay) being present, the traditional term "caries free" is increasingly being replaced with "no obvious decay experience" (Selwitz et al. 2007). Figure 57.1 demonstrates the various thresholds of dental caries diagnosis.

Dental caries is measured via the D_3MF/d_3mf index (upper case letters (DMF) for permanent dentition/lower case (dmf) for deciduous teeth and where the subscript "3" indicates caries at the level into dentine). It is the sum of the number of decayed, missing, or filled teeth (D_3MFT) or decayed, missing, or filled (tooth) surfaces D_3MFS. It was first described by Klein and Palmer (1937). The advantages of the DMF index for measuring dental caries include global acceptance and widespread use for over 60 years, providing an accurate measure of dental caries prevalence. However, it is not without limitations, which include some differences in criteria for caries diagnosis across the globe (e.g., in the USA a sharp probe is used, while in Europe researchers use a blunt probe only to remove plaque prior to inspection

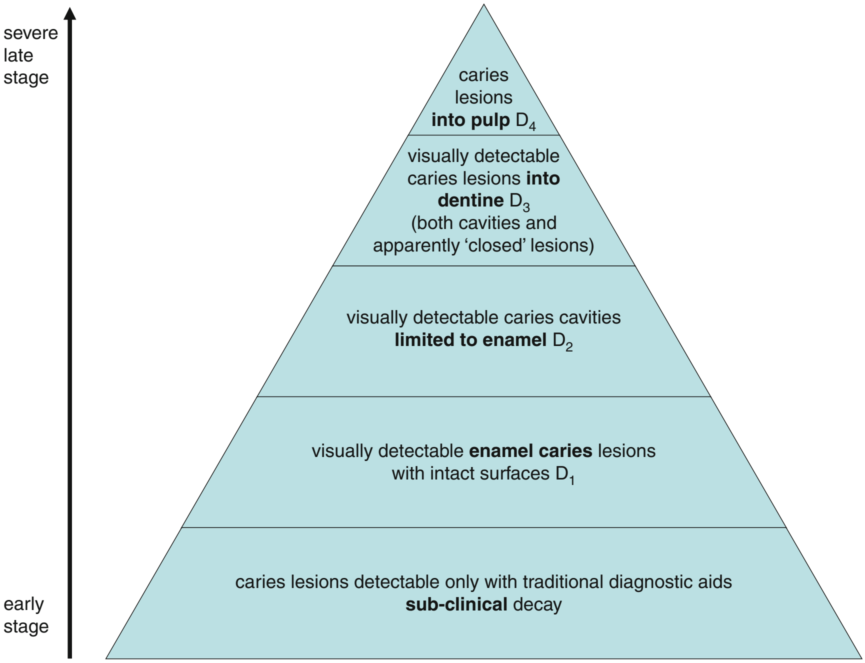

severe
late
stage
early
stage
caries lesions **into pulp** D_4
visually detectable caries lesions **into dentine** D_3 (both cavities and apparently 'closed' lesions)
visually detectable caries cavities **limited to enamel** D_2
visually detectable **enamel caries** lesions with intact surfaces D_1
caries lesions detectable only with traditional diagnostic aids **sub-clinical** decay

Fig. 57.1 Pyramid of thresholds of dental decay (Adapted from Pitts 2001)

of surface); the reason for missing teeth is not always due to caries (e.g., teeth can be lost due to periodontal disease, trauma, extracted as part of orthodontic treatment, or can be congenitally missing); the filled component is influenced by practitioners' decision to restore a tooth, and also new aesthetic restorations can be difficult to detect with a simple visual inspection (Whelton and O'Mullane 2007). The filled (FT) component as a proportion of DMF has been considered the care index – i.e., a measure of the amount of decay which has been treated by restorations ("fillings"). The prevalence of dental caries is usually presented as the mean DMF score or conversely as the proportion of subjects with "free of obvious dental decay experience."

Decay at the tooth surface level (D_3MFS) is a more precise measure of obvious dental caries experience than at the tooth level (D_3MFT) as, for example, a tooth with four or five surfaces affected by caries has the same contribution to the D_3MFT score as a tooth with only one surface decayed. Nevertheless, the D_3MFT remains a useful indicator of the number of teeth that are affected and require treatment. Another way in which caries data are analyzed is on the basis of the type of tooth surface that has been decayed. Thus, differences can be reported between coronal pit and fissure (the biting) surfaces of teeth and the proximal (between teeth) or smooth surface caries (McDonald and Sheiham 1992).

The mean D_3MFT or D_3MFS is usually reported. However, as a significant proportion of many populations have no obvious decay experience and the disease is skewed to a smaller proportion of the population having the majority of the disease, the mean D_3MFT in those with any decay (i.e., $D_3\text{MFT} > 0$) is also used to report caries prevalence.

In older age groups and adults, the use and validity of the D_3MFT is questionable (due to root surface caries in addition to exacerbation of the limitations described above). Other measures of oral health include number of natural teeth present, the absence of all natural teeth – edentulousness, and other measures of the impact of oral health on quality of life are increasingly recognized (Locker 2004).

New systems for recording caries at enamel and dentinal levels have been developed, e.g., the International Caries Detection and Assessment System (ICDAS) – which records decay at a range of levels of severity from early stage enamel to severely decayed lesions involving pulp (Pitts 2004). This system is more complex and requires compressed air to systematically dry the teeth prior to visual inspection. However, it is increasingly being used in epidemiological surveys and clinical trials.

57.2.2 Periodontal Disease

57.2.2.1 Pathogenesis

Periodontal disease is a range of conditions that can be inflammatory, infective, traumatic, metabolic, neoplastic, developmental, and genetic in origin. However, it usually refers to inflammatory disease of the tissues which support the teeth (the periodontium). There are two main conditions: gingivitis and periodontitis.

Table 57.1 Classification of periodontal conditions (Armitage 1999)

I. Gingival diseases
A. Dental plaque-induced gingival diseases
B. Non-plaque-induced gingival lesions
II. Chronic periodontitis
A. Localized
B. Generalized
III. Aggressive periodontitis
A. Localized
B. Generalized
IV. Periodontitis as a manifestation of systemic diseases
V. Necrotizing periodontal diseases
VI. Abscesses of the periodontium
VII. Periodontitis associated with endodontic lesions
VIII. Developmental or acquired deformities and conditions

There is an emerging consensus on the classification of the range of periodontal conditions (Table 57.1). The pathogenesis of periodontal disease is complex and involves the interaction between a pathogenic microflora within the dental biofilm (plaque) on the tooth surface and the host immuno-inflammatory response (Kornman 2008). Gingivitis is a highly prevalent condition, whereby the gingiva (gums) are inflamed. It is reversible by oral hygiene measures. Periodontitis is the more severe and serious condition, where the inflammation extends into the deeper tissues with the loss of supporting connective tissue and the dentoalveolar bone (i.e., the alveolar bone of the jaws that immediately surrounds the teeth). This is known as periodontal attachment loss, and it generally manifests as deepening of the crevice between the gingival tissue and root of the tooth – known as "pocketing." As these pockets deepen, the teeth can become loose, and mastication can be effected, which can lead to tooth loss. It is also possible that gingival recession accompanies the attachment loss and there is no obvious "pocketing."

57.2.2.2 Clinical Diagnosis

The symptoms of gingivitis are limited to mild bleeding gums on toothbrushing, while periodontitis, in its mildest forms, is largely similar. However, as periodontitis progresses, halitosis (i.e., malodors exhaled on breathing; colloquially known as "bad breath") is a common symptom, and when severe, teeth can loosen or even be lost. Both conditions are rarely painful although there is increasing evidence that periodontal disease can lead to a significant impact on oral-health-related quality of life measures (Bernabé and Marcenes 2010).

Diagnosis is based on a clinical examination and radiographs. Using a periodontal probe, clinical probing depths are recorded by measuring the distance from the gingival margin to the depth of the periodontal pocket (the pocket being the space between the tooth and gingivae) using a periodontal probe. The healthy depth is 2 mm; with periodontitis, this depth increases as the periodontium and bone are lost. Initially, a Basic Periodontal Examination (BPE) is undertaken. The full dentition is divided into sextants. The six sextants comprise of the four sets (upper and lower and left and right) of molars and premolars, and the two sets (upper and lower) of canines and incisors. The periodontium surrounding all teeth is probed gently using a ball-ended probe with a colored measurement area corresponding to 3.5–5.5 mm in clinical probing depth and each sextant is scored. Sextants are recorded as a "4" if there is at least one tooth with a pocket greater than 5.5 mm; score "3" for pockets 3.5–5.5 mm; score "2" for no pockets exceeding 3 mm, but calculus (i.e., a hardened form of dental plaque caused by continual accumulation of minerals from saliva on plaque; also known as "tartar") or plaque present; score "1" for no pockets but bleeding on gentle probing; and score "0" for no bleeding or pocketing (Ainamo et al. 1982; Dowell and Chapple 2002). Comprehensive periodontal examination includes measuring the loss of attachment from a fixed point on the tooth and the clinical probing depth at six points around each tooth. Bleeding on probing, the amount of dental plaque and dental calculus (hardened, mineralized dental plaque), horizontal bone loss between the roots, and mobility of the teeth are also recorded. Radiographs are required to demonstrate loss of alveolar bone. The American Academy of Periodontology classifies loss of attachment as mild (1–2 mm of attachment loss), moderate (3–4 mm of attachment loss), and severe (≥5 mm of attachment loss) (Armitage 1999).

57.2.2.3 Epidemiological Measurement

In epidemiology of periodontal disease, cross-sectional surveys are the norm, which usually aim to measure the prevalence and severity of gingivitis and periodontal disease. The epidemiology is complicated by the varying definitions of the disease. The historical perspective that periodontal disease was a single-disease continuum from gingivitis progressing to periodontal disease is now redundant. Separate clinical measures are now used to record gingivitis and periodontitis, although the varied use of different indices further complicates comparison across surveys and populations.

There are a number of indices for measuring gingivitis. The Gingival Index (GI) described by Löe and Silness (1963) gives a score of the level of inflammation from 0 to 3 for the gum at each surface of the teeth based on signs of inflammation. It does not include the pocket depth and therefore the health of the periodontium. More recently, the presence of gingival bleeding upon "gentle probing" has become the standard measure of gingivitis (Greenstein 1984), although the variation and subjectivity of how "gentle" probing is defined and carried out has been questioned (Polson and Caton 1985). Nevertheless, bleeding on probing (BOP) measures are increasingly used in epidemiology; however, there remains significant debate about the discriminatory power, as well as issues related to infection control, and the ethics justifying the inducement of bleeding for epidemiological (not treatment) purposes (Burt and Eklund 2005). A Modified Gingival Index (MGI) has been proposed which does not involve bleeding on probing – but it is recognized that further development of a valid objective measure is still needed (Lobene et al. 1989).

The Community Periodontal Index ofTreatment Needs (CPITN) is frequently used in epidemiological surveys to measure periodontal disease (Ainamo et al. 1982). However, it is increasingly used as a clinical assessment of treatment need or as a screening tool for further periodontal investigation – when it is known as Basic Periodontal Examination (BPE) (Dowell and Chapple 2002). The CPITN measures periodontal pocket depth from a sample of sites in the dentition (rather than measuring the pocket depth from all sites around each tooth in a more detailed periodontal examination). This sampled measure has the inherent problem of potentially underestimating disease prevalence and has been argued as not appropriate for measuring severity or prevalence of periodontal disease (American Academy of Periodontology 1996). Full-mouth recording of the clinical attachment loss and clinical periodontal probing depth are considered more appropriate for recording severity and prevalence of periodontitis. However, this is rarely done due to the fact that it is extremely time-consuming (Burt and Eklund 2005). All partial-mouth recording techniques can result in an underestimation in the prevalence, extent, and severity of disease (Susin et al. 2005). As with gingivitis measures, debate remains around probing force used, and computerized probes have been proposed, although there seems little difference with human scoring (Osborn et al. 1990). Biomarkers are increasingly being sought to measure periodontal disease but, at this stage, are largely experimental.

57.2.3 Tooth Loss

Tooth loss has been considered a form of “dental mortality” (Burt and Eklund 2005). It is a crude measure of general oral health – either as total tooth loss, also known as edentulism, or partially dentate. Furthermore, the World Health Organization defines a functional dentition as one which has at least 20 natural teeth in occlusion; it is also known as the shortened dental arch (Petersen 2003a). Total tooth loss has also been consistently shown to be related to poorer general health (Kandelman et al. 2008).

57.3 Descriptive Epidemiology

57.3.1 Trends in Dental Caries

Global comparisons of the prevalence of dental caries are complicated by variations in diagnostic criteria. Estimates of dental caries prevalence from a large number of countries are collated by the WHO Global Oral Health Database and the WHO Oral Health Country/Area Profile Programme (WHO 2010a). However, as some results are from national surveys with representative samples while others relate only to small local surveys, caution is required in making simplistic intercountry comparisons using the raw data. It is also necessary to understand the public health aims behind the WHO “basic methods” diagnostic criteria employed by most datasets in their databank, and these surveys are only intended to provide an overview of caries prevalence rather than detailed estimates of burden (Petersen 2003b). The prevalence of dental caries varies across the world. The disease is less common and severe in Asia and Latin America and even less common in Africa. Figure 57.2 presents the dental caries levels of 12-year-old children in six global regions as defined by the World Health Organization (WHO) at the turn of the twenty-first century (Petersen 2003b). The WHO also demonstrated that dental caries among 12-year-olds has declined in many developed countries from a high D_3MFT level of around 4.5 in 1980 to about 2.5 in 1998. Similar data across the same period for children in developing countries were lower across the same period, although they have been rising constantly from a D_3MFT of 1.5 to 2.5 (Petersen 2003b).

Marthaler (2004), in a comprehensive review of dental caries trends in industrialized countries from 1953 to 2003, described the emergence of robust dental epidemiological methods and surveys from 1960s. He identified a substantial decline in the prevalence of dental caries from the 1970s and through the 1980s documented over this period, which he related to the introduction of fluoridated toothpaste. However, by 1990, when low prevalence rates had been reached, this decline had begun to slow and had almost stopped. The disease had concentrated in specific teeth and surfaces in both primary and permanent teeth. In permanent teeth, dental caries had not declined on occlusal (biting) surfaces as it had on smooth surfaces (Brown and Selwitz 1995).

Figure 57.3 demonstrates the general trend of caries prevalence from cross-sectional surveys in developed and developing countries during the 1980s and 1990s. These trends can potentially be explained in the context of the risk factors for dental caries (specifically high-sugar diet) and the protective factors (specifically fluoride use). With developing countries shifting to adopt developed (Westernised) culture and diets, sugar consumption has increased, and this has an impact on increasing caries levels. Since the 1960s, affluent-income countries have increasingly adopted caries protective measures – particularly the use of fluoride toothpaste (and in some countries water fluoridation), which has had an effect of reducing caries levels overall. Sales of fluoride toothpaste in the United Kingdom were reported to be less than 5% of total sales in 1970 but had risen to more than 95% of sales by 1977 (Anderson et al. 1981). However, the flattening of the improvement in these countries is perhaps an indication of the limitations in the penetration or reach of fluoride and the underlying unequal distribution of the disease (see below), which may be holding further improvements back.

There are significant inequalities in the distribution of dental caries between and within countries. International data over time are available for 11- to 12-year-olds, which show trends in dental health during a period of nearly 30 years. Figure 57.4 compares the mean DMFT of several European countries for which data were

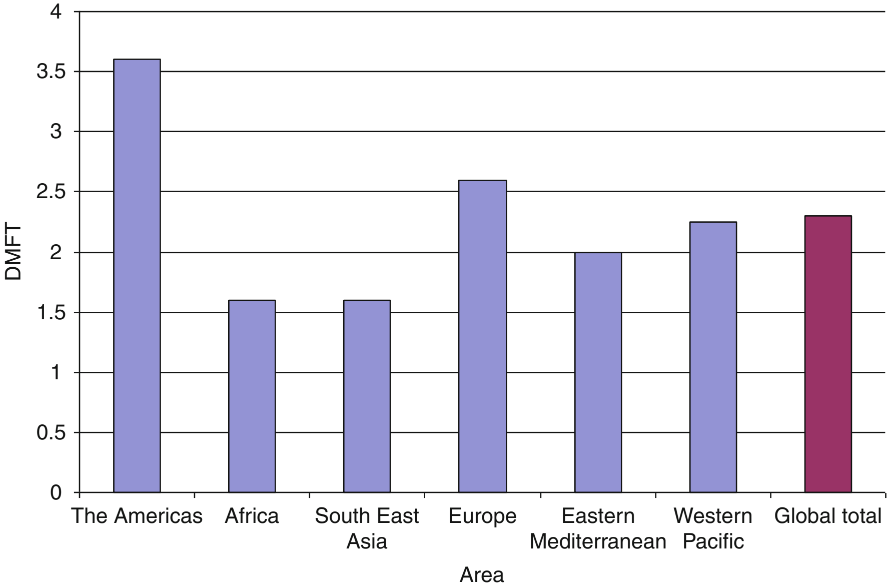


Fig. 57.2 Dental caries experience (D_3 MFT) of 12-year-old children by WHO defined regions, circa 2000 (Data from Petersen 2003b)

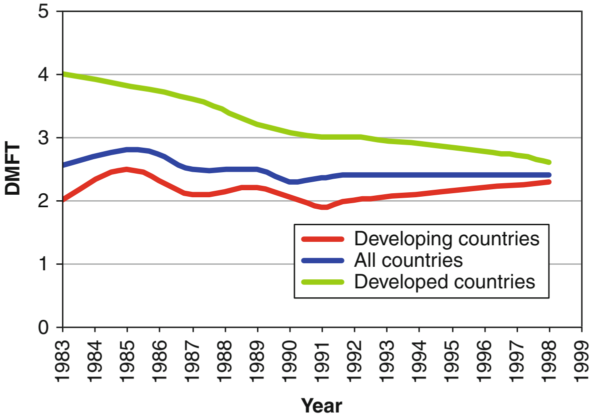


Fig. 57.3 Changing levels of dental caries experience (DMFT) among 12-year-olds in developed, developing, and all countries combined, 1983–1998 (Adapted from Petersen 2003b)

available. Countries in Eastern Europe have consistently higher levels of decay. Overall, it can be seen that for many countries, there was a steep decline in the prevalence of dental caries among 11- to 12-year-olds between 1983 and 1989. This rapid improvement has been followed by a far more gradual improvement since 1989. Marthaler (2004) puts the flatlining of the improvement down to the widening

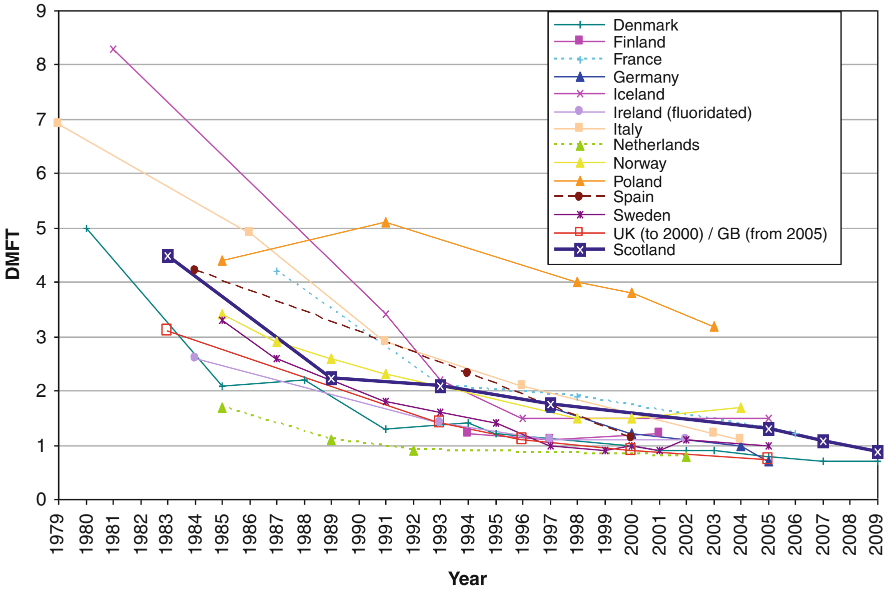


Fig. 57.4 Mean number of decayed, missing and filled teeth (DMFT) in 11- to 12-year-olds, by selected European countries, where data available, 1979–2009 (WHO 2010a)

socioeconomic inequalities which have seen oral health promotion and education approaches reaching the majority of the population but remain out of reach for low socioeconomic groups in society.

Surveys of 5-year-olds perhaps provide a more sensitive barometer of change in the oral health of children, as they include decay in the primary dentition only. Data presented here on the trends in the percentage of 5-year-olds in Scotland with no obvious decay are a typical (if not more comprehensive in terms of methodology and reporting) example of those from developed countries. Comparable data from the national child dental health surveys of the UK Office of Population Census Surveys (OPCS) undertaken in 1983 and 1993, the Scottish Health Boards' Dental Epidemiological Programme (SHBDEP) of surveys between 1987 and 2002, and the National Dental Inspection Programme (NDIP) from 2003 to present are presented in Fig. 57.5. Recent data from NDIP (Merrett et al. 2008; Macpherson et al. 2010a) show a continuation of the marked improvement in the oral health of 5-year-olds in Scotland from 44.6% in 2003 with no obvious decay to 54.1% in 2006 and 64.0% in 2010 – when the (somewhat modest) national target of 60% of children with no obvious decay experience was met. This improvement contrasts favorably with the fairly steady pattern seen over previous years in Scotland, and the considerable investment in child oral health improvement in recent years is a potential explanation.

There are some limited data showing that caries in the primary dentition of 5-year-olds can be used to predict the levels of decay in the permanent dentition

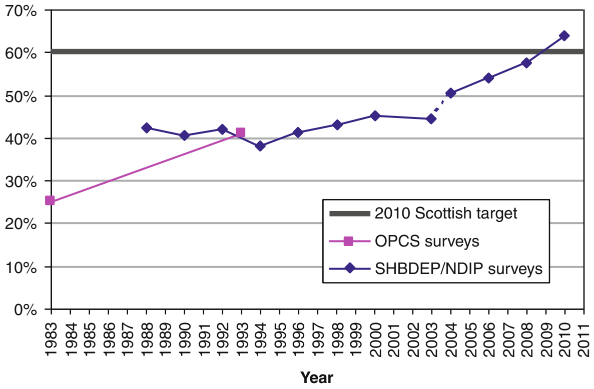


Fig. 57.5 Trends over time in the percentage of 5-year-old children with no obvious decay experience in Scotland, 1983–2010

at 10 years, and in turn caries levels in childhood go onto reflect caries experience in adulthood as reported in a small Norwegian cohort study (Skeie et al. 2006). Susceptibility to the disease continues through adulthood, where it is the major cause of dental pain and tooth loss. Nevertheless, it has been noted that the rate of progression of the disease slows with increasing age (Mejare et al. 2004). There is evidence that dental caries is increasing in prevalence among older adult populations. In the USA, older adults have been found to be getting new caries lesions (root caries) at the same rate as children (Griffin et al. 2004).

57.3.2 Inequalities in Dental Caries

As in childhood caries, adults in vulnerable, high-risk groups have also been identified as having increased dental caries levels. These groups include older adults who are frail (Petersen and Yamamoto 2005) or who live in a residential care home (Chalmers et al. 2005); homeless people (Conte et al. 2006); prisoners (Jones et al. 2004; Treadwell and Formicola 2005); those who are medically compromised (Wyatt and MacEntee 2008) – including those with HIV AIDS (Phelan et al. 2004); recent immigrants (Marthaler 2004); Black and ethnic minority groups (Verrips et al. 2006); and those who are socioeconomically disadvantaged determined by education (Nyyssonen et al. 1984), income (Fukuda et al. 2009), the level of deprivation of the area where they live (Petersen and Yamamoto 2005).

Variations or inequalities in health and disease determined by social or economic determinants have received much attention in terms of policy and research aimed at understanding and trying to tackle these inequalities – e.g., the World Health Organization (Commission on Social Determinants of Health 2008) and the UK government over the past few decades (Bambra et al. 2011). There has been less attention and detailed analysis aimed at understanding between and within country differences and variations in dental caries levels across population groups or areas.

Inequalities in dental caries are observed between different ethnic groups, children from rural and urban areas, and across different area-based socioeconomic circumstances. For example, children from some ethnic backgrounds can

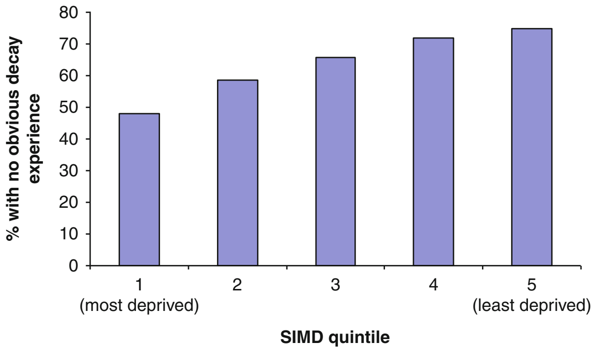


Fig. 57.6 Percentages of 11- to 12-year-old children with no obvious decay experience in permanent teeth by SIMD quintile in the year 2006 in Scotland

have higher caries prevalence than their white contemporaries over and above socioeconomic circumstances (Conway et al. 2007). Children in remote and rural areas of Scotland appear to have better dental health and a higher proportion of filled teeth when compared with those living in cities (Levin et al. 2010).

Figure 57.6 shows that dental decay is more prevalent among children from the most deprived areas in Scotland (Merrett et al. 2009). Area deprivation is measured here by the Scottish Index of Multiple Deprivation (SIMD). SIMD, created by the Scottish Government, for monitoring and planning purposes is calculated using data (from census and other population sources) including six domains of income, employment, housing, health, education, and geographical access to services/telecommunications – which are derived from 31 individual indicators of deprivation at the level of "data zones" neighborhoods. An SIMD score of "1" represents the most deprived areas, and "5" represents the most affluent areas (Scottish Government 2006). These findings are repeated across the developed world (Dye et al. 2010).

Simple measures of inequality like the absolute difference (i.e., the range) and ratio in percentages, from the most deprived to the least deprived, can be calculated from these figures. A range of techniques for quantifying inequalities in health has been developed as described by Kakwani et al. (1997). These analyses include the Gini coefficient, concentration index, slope index of inequalities, relative index of inequality, and population attributable risk. These measures have had limited application in dental data (Antunes et al. 2004; Cheng et al. 2008; Perara and Ekanayake 2008), and while they are not always explicit in terms of the socioeconomic distribution, they generally show a skewed distribution with a small portion of the population (usually the poorest) having the greatest levels of decay.

57.3.3 Trends in Periodontal Disease

Estimating periodontal disease prevalence is complicated by the multiple definitions of the disease and the limited consistency in reporting by age group. There is widely recognized global variation in the prevalence, severity, and rates of disease progression (Löe et al. 1986, 1978; Demmer and Papapanou 2010). According to the World Health Organization, the initial signs of periodontal diseases in adults are highly prevalent across countries, while severe periodontitis is observed in 5–15% of populations (Petersen 2003a).

Recent estimated global variations in reported severe periodontitis given by Demmer and Papapanou (2010) range from 1% among 20- to 29-year-olds in Northern Europe to 39% among those 65 and over in North America. The wide range in these estimates seems to be mainly due to the variation in age ranges of the surveys sampled. Periodontal disease is consistently observed to be more prevalent among men than women, among older than younger age groups, and in black and minority ethnic groups relative to white populations (Demmer and Papapanou 2010). Some studies report a greater prevalence in developing countries (Philstrom et al. 2005), although this is not a consistent finding (Albandar et al. 1999; Demmer and Papapanou 2010) nor is it the case among indigenous or isolated populations (Ronderos et al. 2001).

An earlier WHO report on periodontal health in Europe found that 10% of subjects between 35 and 44 years of age showed advanced periodontal destruction, while moderate disease affected 20–30% of this population. Using CPITN, the prevalence of advanced periodontal disease ranged from 7% to 18%, and moderate disease ranged from 31% to 62% across European countries. The prevalence of periodontitis increases with age, and up to 88% of teeth were affected in older groups (Reich 2001).

In the UK, a survey conducted in 1998 demonstrated that 54% of adults had periodontal disease and 5% had severe periodontitis; prevalence increased with age (Kelly et al. 2000). Oral hygiene was poor in the same population, with nearly three-quarters of dentate adults had visible dental plaque and 73% had calculus present. In the USA, a large survey estimated prevalence of mild periodontitis at 22% of adults and severe periodontitis at 13% (Albandar et al. 1999).

Periodontitis has generally been reported as being rare in younger age groups. The World Oral Health Report of 2003 (Petersen 2003a) reported that globally, while most children have been found to show signs of gingivitis, early onset periodontitis (now referred to as aggressive periodontitis) affected around 2% of children (Albander et al. 1997). In the UK, 35% of children aged 4–18 years had any level of gingivitis or periodontal disease; 40% of 15- to 18-year-olds had gingivitis – with 17% of them having evidence of periodontal disease, which was twice as likely in boys than girls (Gregory et al. 2000). In a cross-sectional study in the United States, more than 11,000 adolescents between 14 and 17 years old were examined; 0.53% had "localized aggressive," and 0.13% had "generalized aggressive" types of periodontal disease. In the same study, Black and Hispanic young people were shown to be at much higher risk for all forms of aggressive periodontitis than white people. Males were more likely to have generalized aggressive periodontitis than females. Moreover, the sex predisposition is different for localized aggressive periodontitis in young people, whereby it was more common in black males than in black females but was more common in white females than in white males (Löe and Brown 1991).

57.3.4 Trends in Tooth Loss (Adults)

As recently as the 1960s/1970s in the UK, loosing a tooth or eventually all teeth – total tooth loss or edentulism – was considered inevitable and dentistry was largely concerned mainly with dental extractions. Anecdotally, a cultural norm in parts of the UK was having all your teeth removed and replaced with shiny full dentures for a common 21st birthday present! (Taylor 1975). Both public expectations and restorative dentistry developed dramatically in the latter parts of the end of the twentieth century and tooth retention has increased particularly in developed countries. Edentulous rates are generally lower in developing countries, although intercountry comparisons are difficult because of the different age groupings employed in national surveys (Petersen 2003a).

Globally, dental caries is considered the principal cause for tooth loss for all ages combined (Oliver and Brown 1993). For those over the age of 45 years, periodontal disease is increasingly recognized as being an important cause of tooth loss (Chestnutt et al. 2000). Some studies suggest it is the main cause of tooth loss (Klock and Haugejordan 1991), while others reaffirm caries is still the most important cause (Chestnutt et al. 2000; Burt and Eklund 2005).

The most recently available data from the UK in 1998 showed that 13% of all adults had no natural teeth, which was projected to decline to 8% by 2008 (Steele et al. 2000). The levels of edentulousness were found to be higher in lower socioeconomic groups and in northern regions and countries of the UK.

The general trends from analysis of edentulous data from Scotland (presented below) are typical (albeit at a higher levels) of the trends observed in developed countries (Burt and Eklund 2005). It should be noted that despite the improvement, the oral health of people in Scotland still lags behind Western European norms (TNS Opinion and Social 2010). The main source of data on the oral health status of adults in Scotland is from the series of UK Adult Dental Health Surveys, which were conducted in 1972, 1978, 1988, and last conducted in 1998 (Kelly et al. 2000). More recently, the Scottish Health Survey (SHeS) from 2008 to 2009 has provided similar data for Scotland. The dental health of adults was poorer in Scotland than in the rest of the UK with recent data from England, Wales, and Northern Ireland showing that the proportion of the population with no natural teeth was 6% in 2009 (Scottish Government 2009; Steele and O'Sullivan 2011). Nevertheless, there has been considerable improvement since in Scotland in recent decades the proportion of adults with no natural teeth declined from 44% in 1972 to 12% in 2008 (Fig. 57.7).

Although Fig. 57.8 shows that as recently as 2008, more than 40% of all people aged over 65 in Scotland had lost all their teeth, there has been a substantial improvement in the oral health of adults since 1972. Moreover, Fig. 57.8 shows very strong cohort effects in the percentage of edentate people. In 1978, 27% of people aged 35–44 were edentate, and by 2008, some 41% of people in the same cohort were edentate. This suggests that there is persistence in oral health over time.

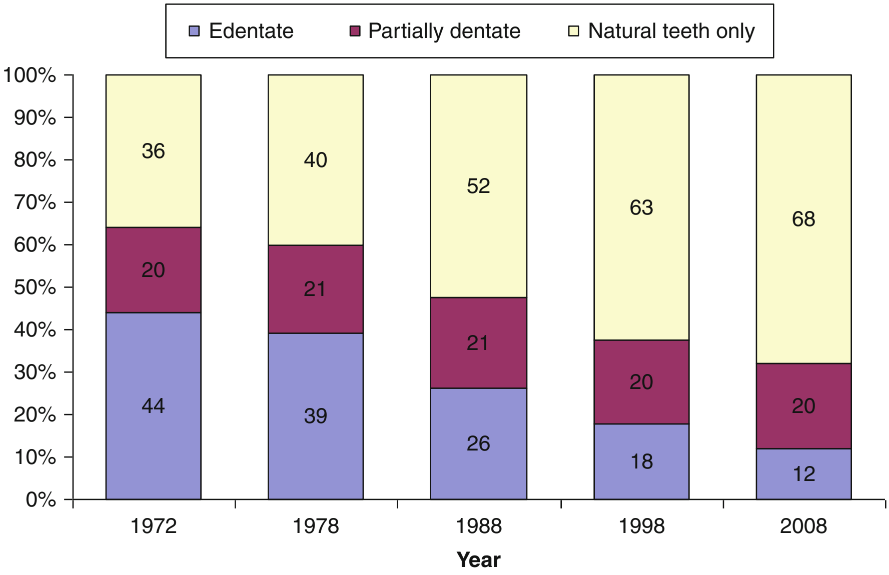


Fig. 57.7 Trends in oral health status of adults in Scotland (1972–2008)

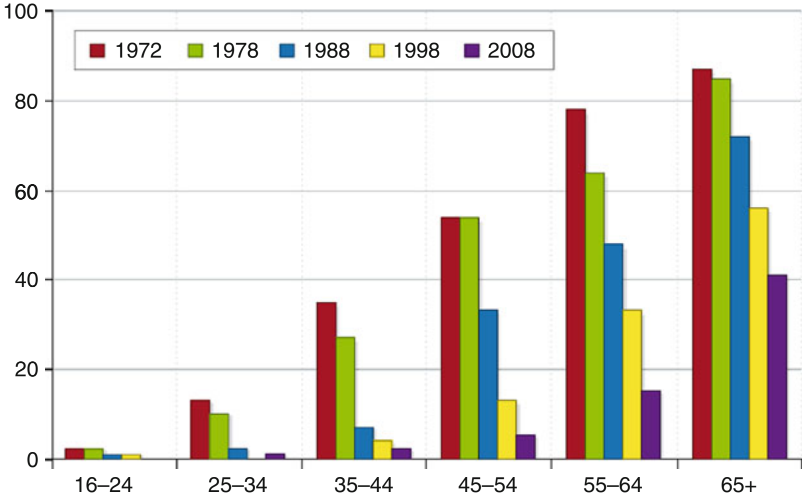
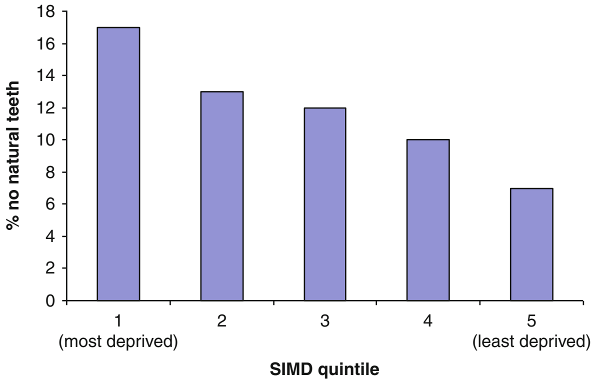


Fig. 57.8 Percentage edentate by age group in Scotland (1972–2008)

Fig. 57.9 Percentage of adults with no natural teeth by Scottish Index of Multiple Deprivation (SIMD) quintile (1998)



While there has been a substantial improvement in adults' oral health in Scotland since 1972, Fig. 57.9 shows that people from the most deprived backgrounds are more likely to be edentate.

Data from the recent Scottish Health Survey (Scottish Government 2009), in conjunction with previous work at the UK level using data from the Adult Dental Health Surveys (Kelly et al. 2000), indicate that the proportion of adults retaining their natural teeth has increased over time. Data from these surveys also suggest that attitudes to dental care have changed significantly during the past 20 years, and an increasing number of people are retaining teeth into old age.

Trends from SHeS and Scottish data from previous the Adult Dental Health Surveys can be applied alongside population projections (General Register Office for Scotland 2009) to predict future numbers of adults with certain dental characteristics. This will provide some insight into the potential demand for dental services. Figure 57.10 shows that the percentage of edentate adults in each age group decreased between 1972 and 2008. It is projected that by 2028, levels of edentulism in the under-55 age groups will have fallen to 1% or less.

57.4 Analytical Epidemiology

Figure 57.11 highlights the potential pathways from determinants to disease and demonstrates the "common risk factors" for oral health and disease (Sheiham and Watt 2000), which recognizes that risk factors are common to many chronic conditions and that there is much overlap between general health and oral health. It is also now widely accepted that poverty or low socioeconomic circumstances are strong determinants for ill health. Similarly, this relationship has regularly been shown in relation to oral diseases. How this relationship is mediated is less well established.

Oral diseases are determined by many of the same risk factors as other conditions (Sheiham and Watt 2000). In summary, (i) diets, particularly those high in sugar

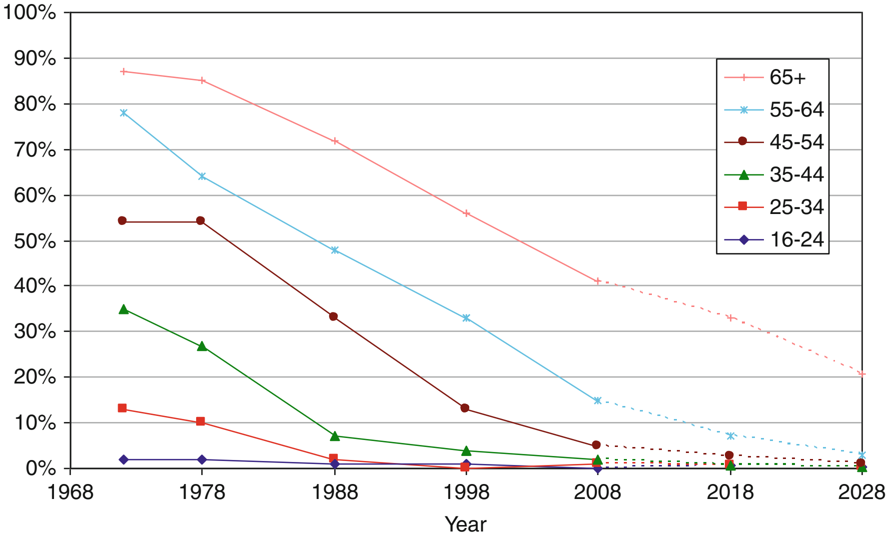


Fig. 57.10 Actual (solid lines) and projected (dotted lines) percentages of edentate adults in Scotland by age group (1972–2028)

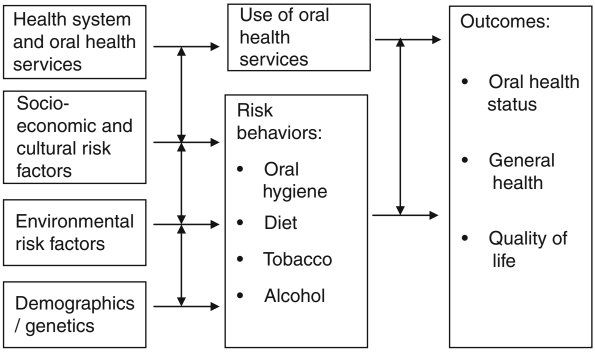


Fig. 57.11 Model of the common risk factor approach (Adapted from Sheiham and Watt 2000; Petersen 2003a)

and/or fat and/or low in fiber and essential vitamins are associated with conditions such as coronary heart disease, stroke, obesity, diabetes, cancers, and dental decay; (ii) smoking is implicated in many diseases, including cancers of the lung, throat, and mouth. In addition, smokers are more likely to have coronary heart disease, diabetes, and gum (periodontal) disease as well as other diseases of the soft tissues of the mouth; (iii) poor oral hygiene is the main cause of gum disease and is also implicated in dental decay, while low general personal hygiene is a factor in many conditions of the skin.

57.4.1 Risk Factors for Dental Caries

This section will outline the risk factors for dental caries. For more details, the reader is referred to comprehensive reviews (Fejerskov and Kidd 2003; Burt and Eklund 2005; Selwitz et al. 2007).

Risk factors for dental caries are multiple, complex, and interrelated; they include bacteria infection, host factors, socioeconomic factors, behavioral or lifestyle factors, and iatrogenic or dental factors. Selwitz and colleagues (2007) consider the behavioral or lifestyle, social, and dental or iatrogenic risk factors most important – however, comprehensive cohort studies have thus far not been undertaken which could unbundle their relative importance. Evidence for each of the key risk factors will be described briefly here.

57.4.1.1 Bacteria

It has long been considered that bacteria are necessary for caries to occur and that caries levels are associated with prevalence of cariogenic bacteria in dental plaque: principally mutans streptococci and lactobacilli (Emilson and Krasse 1985). A systematic review found a strong association between mutans streptococci levels and early childhood caries, although it noted a paucity of longitudinal studies investigating this relationship (Parisotto et al. 2010). The establishment of mutans streptococci colonies in infants has been proposed to be transmitted from mother to child (Berkowitz 2003). However, a direct causal temporal relationship between mutans streptococci and dental caries has yet to be proven – this is partly due to the complexity of the constituents of the oral microflora, with mutans streptococci and lactobacilli also present in healthy mouths. This has led some to propose an imbalance in the composition of the oral microflora (i.e., the interrelationships between the oral mircoorganisms and with their environment) rather than specific bacterial infection. This composition is disrupted and modified (to one with particularly high level of mutans streptococci) by a diet high in refined carbohydrate (Burt and Eklund 2005).

57.4.1.2 Host Factors

Host factors associated with dental caries include deficient salivary flow and quality, tooth morphology and dental enamel quality (determined by levels of mineral – particularly fluoride in the crystalline structure), immunological deficiencies in response to cariogenic bacteria, and genetic factors (Fejerskov and Kidd 2003). A systematic review found some but limited evidence for dental caries genetic variation of host factors including those related to the structure of dental enamel, immunological response to cariogenic bacteria, or the composition of saliva (Shuler 2001).

57.4.1.3 Behavioral or Lifestyle Factors

Behavioral or lifestyle factors include limited access to fluoride, poor oral hygiene, diet with frequent consumption of refined carbohydrates, poor infant feeding and weaning behaviors, and protracted use of sugar-containing oral medicines.

A series of systematic reviews have found that fluoride has a significant caries preventive effect. Risk of dental caries is reduced if fluoride is accessed via, for example, fluoridated toothpaste (Marinho et al. 2003), with increasing preventive effect with fluoride concentrations over 1,000 ppm (Walsh et al. 2010), although there is evidence of increased levels of mild fluorosis if commenced in very young children (Wong et al. 2010). Other clinical caries preventive measures include treatments such as fluoride varnish (Marinho et al. 2002), pit or fissure sealants (Ahovuo-Saloranta et al. 2008) or via fluoridated public drinking water (McDonagh et al. 2000). Similarly, good oral hygiene regimen have been found in a systematic review to be protective for dental caries in children (Harris et al. 2004), although whether this is related to the mechanical dental plaque/bacterial removal versus the fluoride delivery has yet to be fully disentangled.

Historically, in dental epidemiology, etiology and risk has been inferred from descriptive epidemiology studies (rather than analytical case-control and cohort studies). The general relationship between diet (sugars) and dental caries has been established via these methods, including the “causal” role of refined carbohydrates and in particular non-milk extrinsic sugars being widely accepted (Rugg-Gunn and Edgar 1984). This historical evidence has been built up over the past century and includes (i) a series of observational studies of sugar rationing in Norway during World War II (Toverud 1957); (ii) series of observations of the inhabitants of the isolated island of Tristan da Cunha in the South Atlantic as they progressively were exposed to modernized diets as their contact with England increased in the mid-twentieth century (Fisher 1968); (iii) observations of dental caries levels and diet in children in Hopewood House in the 1950s – a care home in Australia – in which children had a diet with restricted sucrose. Comparisons with those of similar age in local population demonstrated dramatic differences in caries prevalence (Sullivan and Harris 1958); and (iv) the infamous Vipehölm experimental study carried out in an institution for adults with learning disabilities and mental health problems in the 1940s and 1950s in Sweden, in which groups were given differing and very high amounts and frequencies of refined sugars in a sticky form. The main findings from this study were that caries risk increased with increased sugar consumption, consumption of sticky form of sugars, and between meal consumption of sticky sugars. They also found that caries risk decreased on removing sticky sugar from the diet and that caries risk is negligible in those with no refined sugar consumption in their diet (Gustafsson et al. 1954).

However, more recent studies have shown that the relationship between dental caries and total sugar consumption or frequency of sugar consumption is less strong than once suspected – particularly in low-caries populations where fluoride exposure is high. A systematic review of observational studies from 1980 to 2000, investigating this relationship, identified studies from countries mainly where fluoride exposure was high and found moderate evidence that sugar was an important risk factor. But the review concluded that limited fluoride exposure was the more important factor (Burt and Pai 2001).

The role of carbonated sugar-sweetened “soft” drinks in dental caries has received some attention not least because of the implication of these drinks in the epidemic of childhood obesity (Mann 2003). A systematic review and meta-analysis investigating the health effects of soft drinks found a small positive association from the two studies with available data (Vartanian et al. 2007). A further systematic review by Kantovitz et al. (2006) found some evidence of a correlation between dental caries and obesity.

A systematic review, albeit of the limited research available, provided inconclusive findings around the relationship between prolonged breast-feeding and early childhood caries, where the confounding effects of socioeconomic status could not be fully taken into account (Valaitis et al. 2000). However, nighttime bottle-feeding and early weaning onto solid foods have been found to increase infant caries risk (Douglass et al. 2001).

A review of the literature for the development of clinical prevention guidelines in Scotland concluded that there was some, albeit limited, evidence supporting increased risk of dental caries following long-term use of sugar-sweetened medicines (Scottish Intercollegiate Guidelines Network 2000).

57.4.1.4 Social Factors

Social factors associated with dental caries risk include individual low income or poverty, low educational attainment, low occupational social class, as well as deprivation associated with area of residence, and in some countries, lack of dental insurance coverage (Gratix and Holloway 1994; Sweeney et al. 1999). A recent life-course study analyzing social mobility over time from a birth cohort in New Zealand identified increased adult caries risk associated with low socioeconomic status in childhood into adulthood, compared to persistent high socioeconomic status and that downward social mobility increased risk, while upward social mobility decreased risk (Thomson et al. 2004). The effect of low socioeconomic status has been proposed as an independent risk factor beyond known explanatory behavioral factors in a Finnish analysis (Millen 1987). Understanding the pathway between low socioeconomic status and dental caries is yet to be fully explored.

57.4.1.5 Dental or Iatrogenic Factors

Dental or iatrogenic factors which can increase caries risk include poor dental restorations or fillings, poor partial dentures, orthodontic appliances, and limited use of (protective) fissure sealants (Fejerskov and Kidd 2003; Benson et al. 2005). A recent Cochrane review found that fissure sealants were effective at preventing dental caries particularly in children’s molar teeth (Ahovuo-Saloranta et al. 2008).

57.4.2 Risk Factors for Periodontal Disease

In this section, the current knowledge about risk factors for gingivitis and periodontal disease is briefly summarized. For more detail, the reader is referred to comprehensive reviews (Philstrom et al. 2005; Demmer and Papapanou 2010; Stabholz et al. 2010).

Poor oral hygiene and supragingival dental plaque have been shown in experimental studies to have a causal relationship with gingivitis (Löe et al. 1965). There does not seem to be any evidence of specific microorganisms responsible, and the condition is considered a non-specific infection (Philstrom et al. 2005).

There is strong evidence from observational epidemiology studies and systematic reviews of these studies that the risk factors associated with increased periodontal disease prevalence include (i) oral microorganisms – both nature and quantity of oral microflora, with a wide range of microorganisms implicated (Haffajee et al. 2004); (ii) dental plaque and calculus which harbor microorganisms, and (iii) poor oral hygiene (Lang et al. 1973; Philstrom et al. 2005). Furthermore, there is increasing evidence of increased periodontal risk associated with genetic factors – particularly variations in genes associated with the inflammatory response (Michalowicz et al. 2000; Stabholz et al. 2010); tobacco smoking (Klinge and Norlund 2005); diabetes mellitus (Ryan et al. 2003); HIV infection and AIDS – particularly in high incidence populations such as South Africa (Scheutz et al. 1997; Arendorf et al. 1998); conditions which impair immune response (Philstrom et al. 2005); low socioeconomic status (Sheiham and Nicolau 2005); psychosocial factors (Nicolau et al. 2007a); and obesity (Chaffee and Weston 2010).

Systematic reviews also reveal that there is more moderate level or limited evidence for periodontal disease risk being associated with the following factors: heavy alcohol consumption (Tezal et al. 2004), nutritional deficiencies (van der Putten et al. 2009), osteoporosis and hormone replacement therapy (Philstrom et al. 2005), and psychosocial stress (Peruzzo et al. 2007). Men are consistently at more risk than women for destructive periodontal disease, and sex hormones have been suggested as having a role, although gender differences in oral hygiene and attendance patterns are also possible explanatory variables (Kelly et al. 2000; Shiau and Reynolds 2010). Dental treatment that predisposes to plaque build-up or oral malocclusions may also be contributory factors (Bollen 2008).

Conversely, there is also moderate level evidence that periodontal diseases are implicated as risk factors in the etiological pathway of other diseases and conditions including adverse pregnancy outcomes – particularly preterm birth and low birth weight (Xiong et al. 2006), cardiovascular disease (Humphrey et al. 2008), stroke (Scannapieco et al. 2003a), pulmonary disease (Scannapieco et al. 2003b; Azarpazhooh and Leake 2006), and type 2 diabetes mellitus (Demmer et al. 2008). The cause and effect relationship between periodontal disease and diabetes is still not clear, and a bidirectional or two-way interrelationship has been proposed. It has been reported that periodontal disease can potentiate the development of diabetes and interfere with glycemic control through pro-inflammatory pathways; however, it has also been shown that periodontitis is itself a complication of diabetes (Iacopino 2001).

57.5 Prevention and Dental Public Health Approaches

As with all epidemiological study of disease, its ultimate aim is prevention.

The WHO Ottawa Charter for Health Promotion (WHO 1986) provides a useful structure with which dental caries can be prevented at a population level. The charter includes developing public health policy, reorientating health services, creating supportive environments, and working with communities, as well as personal skills for behavior change.

57.5.1 Dental Caries

There has been a raft of clinical guidelines on prevention for dental caries produced recently (e.g., Scottish Intercollegiate Guidelines Network 2000; Scottish Intercollegiate Guideline Network 2005; European Academy of Paediatric Dentistry 2009; Scottish Dental Clinical Effectiveness Programme 2010; American Dental Association 2011). These documents focus on grading and interpreting the epidemiology and clinical trial evidence into practical advice and recommendations for clinical practitioners. The role of health services to deliver prevention activities is important, but oral health promotion and improvement of activities in other settings and with other partners are also essential.

It is widely acknowledged that for the prevention of dental caries, these approaches need to focus on activities that will reduce the frequency and amount of sugars in the diet and on programs that will deliver fluoride and fissure sealants. Influencing diet remains a difficult challenge both at the population level and the individual patient level. At the population level, food consumption and choice are determined by many factors, not least access, affordability, culture, education, cooking skills, food regulation, and labeling. Changing dietary behavior at the population level requires concerted multidisciplinary working with a wide range of relevant organizations, agencies, policy makers, businesses, and professionals, as well as consumers. Professional dietary advice in dental practices is usually based on providing one-to-one dietary advice based on individualized health education models. However, there is evidence from a systematic review in a dental setting that this education approach is largely ineffective (Sprod et al. 1996).

Population programs involving the delivery of fluoride have been shown to be effective with a range of different interventions. These programs have included community water fluoridation, school water fluoridation, fluoridated salt, school supervised toothbrushing, professional topical fluoride varnish or gel application in schools, mouth rinsing in schools, provision of fluoride tablets in schools, and fissure sealant programs in school children (Selwitz et al. 2007). There is some concern expressed about the safety of fluoride. However, a systematic review in relation to water fluoridation by McDonagh and colleagues (2000) has shown that the only adverse effect is mild dental fluorosis (i.e., mottling) – although the quality of studies was described as poor. They also identified a need for more longitudinal follow-up studies at the tooth surface level including different socioeconomic subgroups.

57.5.2 Periodontal Disease

Prevention of chronic gingivitis and periodontitis is based on the control of the key risk factor, namely, the buildup of the oral biofilm (dental plaque) which forms on the teeth in the absence of effective oral hygiene (toothbrushing). However, it is also important to take into account the various additional risk factors including smoking, socioeconomic factors, and diabetes control in the prevention and treatment planning of periodontal diseases. It should be noted there are big differences in the potential for control and prevention between gingivitis, chronic periodontitis, and aggressive periodontitis. Gingivitis and chronic periodontitis can be prevented by both professional and personal good oral hygiene regimen. Patients who suffer from aggressive forms of the disease may be hyperresponsive to relatively little plaque. Periodontal susceptibility tests are not yet applicable to clinical practice, and therefore good oral hygiene is the only preventative measure available.

Löe and colleagues (1965) demonstrated that within 24 hours of cessation of toothbrushing, the oral biofilm begins to develop on the teeth leading to gingivitis in 2–3 weeks, which is reversed within 1 week if toothbrushing returns. While the destructive effects of chronic periodontitis cannot be reversed, it has also been shown that professionally administered oral hygiene (one-to-one in a clinical setting) can slow or stop periodontitis, and tooth loss can be avoided (Axelsson et al. 2004).

Community and school-based health education promotion programs have been found to be effective in reducing dental plaque and gingivitis levels in the short-term (6 months), but not in the long-term or in relation to periodontal disease (Watt and Marinho 2005). Moreover, school-based oral health education approaches have been found to widen oral health inequalities, with those from higher socioeconomic groups more able, or to take on health messages and behaviors (Schou and Wight 1994). Nevertheless, the WHO oral health promotion strategy proposed for low-income countries with high-prevalence periodontal disease and limited dental services and public health programs aimed at educating and promoting oral hygiene practices would be cost-effective (Petersen 2003a). They also suggested that oral health promotion in general needs to be integrated to the wider public health agenda and address the socioeconomic and environmental, as well as the behavioral, risk factors of oral disease. In this regard, tobacco control should be an important aspect of preventive programs for periodontal disease as part of a common risk factor approach to health – as smoking is a risk factor for many chronic diseases (Sheiham and Watt 2000).

57.5.3 Dental Public Health

Dental public health approaches are concerned with implementing interventions and activities aimed at preventing oral diseases. This work is often more positively framed as health improvement or health promotion, and it needs to be done via a strategic approach based on appropriate assessment of the population/community needs alongside the resources and assets available. Health promotion cannot be done within the boundaries of health services alone, rather by working in partnerships such as those between health services, local government organizations, education settings, private sector/business, local organizations, communities, and people.

An example of this approach is the *Childsmile* national (country-wide) child oral health improvement program for Scotland (which the authors of this chapter are heavily involved with). It is a comprehensive health promotion intervention, including community development activities and service redesign as major components (Macpherson et al. 2010b; Turner et al. 2010). It is not simply a dental health education program, whereby oral health advice is passively given – it includes "active" interventions such as nursery (kindergarten) and school toothbrushing alongside healthy snack policies; distribution of free toothbrushes and toothpaste from primary care health services and in nursery schools; referral to dental services from community public health nurses (health visitors) from birth; a targeted nursery school fluoride varnish program; and dental services reoriented to focus on prevention interventions in children such as fluoride varnish, toothbrush demonstration, and tailored dietary advice delivered by other members of the dental team including dental nurses – not only dentists. One of the key challenges this program faces is in tackling oral health inequalities (described via the epidemiology in Sect. 57.3.2 above) and in deciding whether this is best done in a targeted approach focusing efforts on those at greatest need from deprived backgrounds and communities or in a universal, population-wide approach. This has been debated at length for *Childsmile* and resolved, in terms of the strategy, via a combination of the two approaches (Shaw et al. 2009). This approach follows the "Marmot" strategy for tackling health inequalities globally (Marmot et al. 2008) and in England (Marmot et al. 2010) described as "proportionate universalism," which recognizes that inequalities are a gradient and not just extremes of health differences between rich and poor. A major evaluation program has been established to assess the effectiveness of *Childsmile* and its components (Turner et al. 2010). The national epidemiology program of surveys of dental caries in children will be central to the evaluation and will be used to assess the impact on oral health improvement in terms of reduced dental caries and on whether oral health inequalities in terms of the socioeconomic distribution of children's dental caries has reduced. The first paper has recently been published demonstrating a clear improvement in oral health in terms of reduction of caries in 5-year-olds in Scotland, following the introduction of a national supervised toothbrushing programme in nursery schools for 3- and 4-year-olds (Macpherson et al. 2013)

57.6 Conclusions

Dental epidemiology is focused on better understanding the burden and causes of oral diseases. It is increasingly used to evaluate the effectiveness of dental public health programs aimed at improving population oral health (e.g., Turner et al. 2010; McMahon et al. 2011).

The future of dental epidemiology has a number of outstanding issues. With regard to dental caries descriptive epidemiology, the epidemiological threshold for dental caries diagnosis is an important consideration in measuring the prevalence of disease, and there are strong arguments to begin to look at enamel level lesions or subclinical levels. Relatively recent proposals also include a new index, the Significant Caries Index (SiC), which aims to bring attention to the individuals with the highest caries values in each population under investigation. The SiC is calculated as follows: (i) individuals are ranked by their DMFT scores, (ii) the one-third of the population with the highest caries scores is selected and (iii) the mean DMFT for this group is calculated (Bratthall 2000). This index focuses attention on children with the greatest amount of dental caries, and the WHO has begun to adopt the SiC as part of their country level targets for 12-year-olds; thus, countries should aim to have a DMFT of 3 or less for the whole population, and then they should aim to have a SiC of 3 or less DMFT for the one-third of children with the highest levels of caries (WHO 2010b).

The limited understanding of the nature and extent of inequalities in dental caries demonstrates a need to undertake detailed analysis of the trends associated with inequalities in distribution from survey data. These descriptive approaches need to be complemented with analytical epidemiology involving new, large cohort studies to investigate fresh the risk factors for dental caries from the social determinants to the microbiological and genetic levels but also including better quantification of diet and fluoride use. This could include social epidemiology approaches set out in Berkman and Kawachi (2000), life-course epidemiology (Kuh and Ben-Shlomo 2004; Nicolau et al. 2007a), and genetic epidemiology (Burton et al. 2005).

Descriptive epidemiology of periodontal disease needs to be developed conforming to standard definitions and classifications of disease. As with dental caries, the etiological analytical epidemiology literature is somewhat limited in terms of the quantity and quality of longitudinal studies which could temporally determine causal risk factors in the etiology of periodontal disease. Nicolau et al. (2007b) propose the use of life-course epidemiological approaches to investigate oral disease etiology. This approach has the potential to better elucidating the social, ecological, temporal, behavioral, and biological or genetic disease pathways and mechanisms.

References

Ahovuo-Saloranta A, Hiiri A, Nordblad A, Mäkelä M, Worthington HV (2008) Pit and fissure sealants for preventing dental decay in the permanent teeth of children and adolescents. Cochrane Database Syst Rev 4:CD001830. doi:10.1002/14651858.CD001830.pub3

Ainamo J, Barmes D, Beagrie G, Cutress T, Martin J, Sardo-Infirri J (1982) Development of the World Health Organisation (WHO) Community Periodontal Index of Treatment Needs (CPITN). Int Dent J 32:281–291

Albander JM, Brown LJ, Löe H (1997) Clinical features of early-onset periodontitis. J Am Dent Assoc 128:1393–1399

Albandar JM, Brunelle JA, Kingman A (1999) Destructive periodontal disease in adults 30 years of age and older in the United States, 1988–1994. J Periodontol 70:13–29

American Academy of Periodontology (1996) Consensus report: periodontal diseases: epidemiology and diagnosis. Ann Periodontol 1:1

American Dental Association (2011) Center for evidence-based dentistry. ADA clinical recommendations. http://ebd.ada.org/ClinicalRecommendations.aspx. Accessed Oct 2011

Anderson RJ, Bradnock G, James PM (1981) The changes in the dental health of 12-year-old school children in Shropshire. A review after an interval of 10 years. Br Dent J 150:278–281

Antunes JLF, Narvai PC, Nugent ZJ (2004) Measuring inequalities in the distribution of dental caries. Community Dent Oral Epidemiol 32:41–48

Arendorf TM, Bredekamp B, Cloete CA, Sauer G (1998) Oral manifestations of HIV infection in 600 South African patients. J Oral Pathol Med 27:176–179

Armitage GC (1999) Development of a classification system for periodontal diseases and conditions. Ann Periodontol 4:1–6

Axelsson P, Nystrom B, Lindhe J (2004) The long-term effect of a plaque control program on tooth mortality, caries and periodontal disease in adults. Results after 30 years of maintenance. J Clin Periodontol 31:749–757

Azarpazhooh A, Leake JL (2006) Systematic review of the association between respiratory diseases and oral health. J Periodontol 77:1465–1482

Bambra C, Smith KE, Garthwaite K, Joyce KE, Hunter DJ (2011) A labour of Sisyphus? Public policy and health inequalities research from the Black and Acheson Reports to the Marmot Review. J Epidemiol Community Health 65:399–406

Benson PE, Shah AA, Millett DT, Dyer F, Parkin N, Vine RS (2005) Fluorides, orthodontics and demineralization: a systematic review. J Orthod 32:102–114

Berkman LF, Kawachi I (eds) (2000) Social epidemiology. Oxford University Press, Oxford

Berkowitz RJ (2003) Acquisition and transmission of mutans streptococci. J Calif Dent Assoc 31:135–138

Bernabé E, Marcenes W (2010) Periodontal disease and quality of life in British adults. J Clin Periodontol 7:968–972

Bollen AM (2008) Effects of malocclusions and orthodontics on periodontal health: evidence from a systematic review. J Dent Educ 72:912–918

Bratthall D (2000) Introducing the Significant Caries Index together with a proposal for a new global oral health goal for 12-year-olds. Int Dent J 50:378–384

Brown LJ, Selwitz RH (1995) The impact of recent changes in the epidemiology of dental caries on guidelines for the use of dental sealants. J Public Health Dent 55:274–291

Burt BA, Eklund SA (eds) (2005) Dentistry, dental practice and the community. Elsevier Saunders, St Louis, pp 203–209

Burt BA, Pai S (2001) Sugar consumption and caries risk. A systematic review. J Dent Res 65:1154–1158

Burton PR, Tobin MD, Hopper JL (2005) Key concepts in genetic epidemiology. Lancet 366: 941–951

Chaffee BW, Weston SJ (2010). The association between chronic periodontal disease and obesity: a systematic review with meta-analysis. J Periodontol 81:1708–1724

Chalmers JM, Carter KD, Spencer AJ (2005) Caries incidence and increments in Adelaide nursing home residents. Spec Care Dentist 25:96–105

Cheng NF, Han PZ, Gansky SA (2008) Methods and software for estimating health disparities: the case of children's oral health. Am J Epidemiol 168:906–914

Chestnutt I, Binnie V, Taylor M (2000) Reasons for tooth extraction in Scotland. J Dent 28:295–297

Commission on Social Determinants of Health (2008) CSDH final report: closing the gap in a generation: health equity through action on the social determinants of health. World Health Organization, Geneva

Conte M, Broder HL, Jenkins G, Reed R, Janal MN (2006) Oral health, related behaviors and oral health impacts among homeless adults. J Public Health Dent 66:276–278

Conway DI, Quarrell I, McCall DR, Gilmour H, Bedi R, Macpherson LM (2007) Dental caries in 5-year-old children attending multi-ethnic schools in Greater Glasgow – the impact of ethnic background and levels of deprivation. Community Dent Health 24:161–165

Demmer RT, Papapanou PN (2010) Epidemiologic patterns of chronic and aggressive periodontitis. Periodontology 2000 53:28–44

Demmer RT, Jacobs DR Jr, Desvarieux M (2008) Periodontal disease and incident type 2 diabetes: results from the First National Health and Nutrition Examination Survey and its epidemiologic follow-up study. Diabetes Care 31:1373–1379

Douglass JM, Rinanoff N, Tang JM, Altman DS (2001) Dental caries patterns and oral health behaviours in Arizona infants and toddlers. Community Dent Oral Epidemiol 29:14–22

Dowell P, Chapple IL (2002) The British Society of Periodontology referral policy and parameters of care. Dent Update 29:352–353

Dye BA, Arevalo O, Vargas CM (2010) Trends in paediatric dental caries by poverty status in the United States, 1988–1994 and 1999–2004. Int J Paediatr Dent 20:132–143

Emilson CG, Krasse B (1985) Support for and implications for the specific plaque hypothesis. Scand J Dent Res 93:96–104

European Academy of Paediatric Dentistry (2009) Guidelines on the use of fluoride in children: an EAPD policy document. Eur Arch Paediatr Dent 10:129–134

Fejerskov O (1997) Concepts of dental caries and their consequences for understanding the disease. Community Dent Oral Epidemiol 25:5–12

Fejerskov O, Kidd EAM (eds) (2003) Dental caries: the disease and its clinical management. Blackwell Monksgaard, Copenhagen

Fisher FJ (1968) A field survey of dental caries, periodontal disease and enamel defects in Tristan da Cunha. Br Dent J 125:447–453

Fukuda H, Kuroda K, Ohsaka T, Takatorige T, Nakura I, Saito T (2009) Oral health status among low-income people admitted to Osaka Socio-Medical Center in Japan. Int Dent J 59: 96–102

General Register Office for Scotland (2009) Projected population of Scotland (2008-based). General Register Office for Scotland, Edinburgh

Gratix D, Holloway PJ (1994) Factors of deprivation associated with dental caries in young children. Community Dent Health 11:66–70

Gregory J, Lowe S, Bates CJ, Prentice A, Jackson LV, Smithers G, Wenlock R, Farron M (2000) National Diet and Nutrition Survey: young people aged 4 to 18 years, vol 1. Report of the diet and nutrition survey. The Stationery Office, London

Greenstein G (1984) The role of bleeding on the diagnosis of periodontal disease. J Periodontol 55:648–648

Griffin SO, Griffin PM, Swann JL, Zlobin N (2004) Estimating rates of new root caries in older adults. J Dent Res 83:634–638

Gustafsson BE, Quensel CE, Lanke LS, Lundquist C, Grahnen H, Bonow BE, Krasse B (1954) The Vipeholm dental caries study; the effect of different levels of carbohydrate intake on caries activity in 436 individuals observed for five years. Acta Odontol Scand 11:232–364

Haffajee AD, Bogren A, Hasturk H, Feres M, Lopez NJ, Socransky SS (2004) Subgingival microbiota of chronic periodontitis subjects from different geographic locations. J Clin Periodontol 31:996–1002

Harris R, Nicoll AD, Adair PM, Pine CM (2004) Risk factors for dental caries in young children: a systematic review of the literature. Community Dent Health 21(1S):71–85

Humphrey LL, Fu R, Buckley DI, Freeman M, Helfand M (2008) Periodontal disease and coronary heart disease incidence: a systematic review and meta-analysis. JGIM 23:2079–2086

Iacopino AM (2001) Periodontitis and diabetes interrelationships: role of inflammation. Ann Periodontol 6:125–137

Ismail A (2004) Diagnostic levels in dental public health planning. Caries Res 38:199–203

James B, Robbins WJ, Schwartz RS (2001) Fundamentals of operative dentistry: a contemporary approach, 2nd edn. Quintessence, Carol Stream

Jones CM, McCann M, Nugent ZJ (2004) Scottish Prisons' Dental Health Survey 2002. Scottish Executive, Edinburgh. http://www.scotland.gov.uk/Publications/2004/02/18868/32855. Accessed Oct 2011

Kakwani N, Wagstaff A, Vandoorslaer E (1997) Socioeconomic inequalities in health: measurement, computation, and statistical inference. J Econom 77:87–103

Kandelman D, Petersen PE, Ueda H (2008) Oral health, general health, and quality of life in older people. Spec Care Dentist 28:224–236

Kantovitz KR, Pascon FM, Rontani RM, Gavião MB (2006) Obesity and dental caries – a systematic review. Oral Health Prev Dent 4:137–144

Kelly M, Steele J, Nuttal N, Bradnock G, Morris J, Nunn J, Pine C, Pitts N, Treasure E, White D (2000) Adult Dental Health Survey: oral health in the United Kingdom in 1998. The Stationery Office, London

Kidd EAM (1987) Essentials of dental caries: the disease and its management. Wright, London

Kidd EA, Fejerskov O (2004) What constitutes dental caries? Histopathology of carious enamel and dentin related to the action of cariogenic biofilms. J Dent Res 83:C35-C38

Klein H, Palmer CE (1937) Dental caries in American Indian children, Public Health Bulletin 239. US Government Printing Office, Washington DC

Klinge B, Norlund A (2005) A socio-economic perspective on periodontal diseases: a systematic review. J Clin Periodontol 32(s6):314–325

Klock K, Haugejordan O (1991) Primary reasons for extraction of permanent teeth in Norway: changes from 1968 to 1998. Community Dent Oral Epidemiol 9:336–341

Kornman KS (2008) Mapping the pathogenesis of periodontitis: a new look. J Periodontol 79(s8):1560–1568

Kuh D, Ben-Shlomo Y (eds) (2004) A life course approach to chronic disease epidemiology, 2nd edn. Oxford University Press, Oxford

Lang NP, Cumming BR, Löe H (1973) Toothbrushing frequency as it relates to plaque development and gingival health. J Periodontol 44:396–405

Levin KA, Davies CA, Douglas GV, Pitts NB (2010) Urban-rural differences in dental caries of 5-year old children in Scotland. Soc Sci Med 71:2020–2027

Lobene RR, Mankodi SM, Ciancio SG, Lamm RA, Charles CH, Ross NM (1989) Correlation among gingival indices: a methodology study. J Periodontol 60:159–162

Locker D (2004) Oral health and quality of life. Oral Health Prev Dent 2(Suppl 1):247–253

Löe H, Brown LJ (1991) Early onset periodontitis in the United States of America. J Periodontol 62:608–616

Löe H, Silness J (1963) Periodontal disease in pregnancy. I Prevalence and severity. Acta Odontol Scand 21:533–551

Löe H, Anerud A, Boysen H, Smith M (1978) The natural history of periodontal disease in man. The rate of periodontal destruction before 40 years of age. J Periodontol 49:607–620

Löe H, Theilade E, Jensen SB (1965) Experimental gingivitis in man. J Periodontol 36:177–187

Löe H, Anerud A, Boysen H, Morrison E (1986) Natural history of periodontal disease in man. Rapid, moderate and no loss of attachment in Sri Lankan laborers 14 to 46 years of age. J Clin Periodontol 13:431–445

Macpherson LM, Anopa Y, Conway DI, McMahon AD (2013) National supervised toothbrushing program and dental decay in Scotland. J Dent Res 92:109–13

Macpherson LMD, Conway DI, Goold S, Jones CM, McCall DR, Merrett MCW, Pitts NB (2010a) National Dental Inspection Programme of Scotland. Scottish Dental Epidemiological Co-ordinating Committee, Dundee. http://www.scottishdental.org/index.aspx?o=2153. Accessed Oct 2011

Macpherson LMD, Ball GE, Brewster L, Duane B, Hodges CL, Wright W, Gnich W, Rodgers J, McCall DR, Turner S, Conway DI (2010b) Childsmile: the national child oral health improvement programme in Scotland. Part 1: Establishment and development. Br Dent J 209:73–78

Mann J (2003) Sugar revisited – again. Bull World Health Organ 81:552

Marinho VCC, Higgins JPT, Logan S, Sheiham A (2002) Fluoride varnishes for preventing dental caries in children and adolescents. Cochrane Database Syst Rev 3:CD002279. doi:10.1002/14651858.CD002279

Marinho VCC, Higgins JPT, Logan S, Sheiham A (2003) Fluoride toothpastes for preventing dental caries in children and adolescents. Cochrane Database Syst Rev 1:CD002278. doi:10.1002/14651858.CD002278

Marmot MG, Freil S, Bell R, Houweling TAJ, Taylor S, on behalf of the Commission on Social Determinants of Health (2008) Closing the gap in a generation: health equity through action on the social determinants of health. Lancet 372:1661–1669

Marmot MG, Allen J, Goldblatt P, Boyce T, McNeish D, Grady M, Geddes I, on behalf of the Marmot Review (2010) Fair society, healthy lives: Strategic review of health inequalities in England post-2010. The Marmot Review, London

Marthaler TM (2004) Changes in dental caries 1953–2003. Caries Res 38:173–181

McDonagh MS, Whiting PF, Wilson PM, Sutton AJ, Chestnutt I, Cooper J, Misson K, Bradley M, Treasure E, Kleijnen J (2000) Systematic review of water fluoridation. BMJ 321:855–859

McDonald SP, Sheiham A (1992) The distribution of caries on different tooth surfaces at varying levels of caries – a compilation of data from 18 previous studies. Community Dent Health 9:39–48

McMahon AD, Blair Y, McCall DR, Macpherson LM (2010) The dental health of three-year-old children in Greater Glasgow, Scotland. Br Dent J 209:E5

McMahon AD, Blair Y, McCall DR, Macpherson LM (2011) Reductions in dental decay in 3-year old children in Greater Glasgow and Clyde: Repeated population inspection studies over four years. BMC Oral Health 11:29

Mejare I, Stenlund H, Zelezny-Homlund C (2004) Caries incidence and lesion progression from adolescence to young adulthood: a prospective 15-year cohort study in Sweden. Caries Res 38:130–141

Merrett MC, Goold S, Jones CM, McCall D, Macpherson LMD, Nugent ZJ, Topping GV (2008) National Dental Inspection Programme of Scotland. Report of the 2008 Survey of P1 Children. Scottish Dental Epidemiology Coordinating Committee, Dundee. http://www.scottishdental.org/index.aspx?o=2153. Accessed Oct 2011

Merrett MCW, Conway DI, Goold S, Jones CM, McCall DR, McMahon AD, Macpherson LMD, Pitts NB (2009) National Dental Inspection Programme of Scotland 2009. Report of the 2009 Survey of P7 Children. Scottish Dental Epidemiological Coordinating Committee, Dundee. http://www.scottishdental.org/index.aspx?o=2153. Accessed Oct 2011

Michalowicz BS, Diehl SR, Gunsolley JC, Sparks BS, Brooks CN, Koertge TE, Califano JV, Burmeister JA, Schenkein HA (2000) Evidence of a substantial genetic basis for risk of adult periodontitis. J Periodontol 71:1699–1707

Millen A (1987) Role of social class in caries occurrence in primary teeth. Int J Epidemiol 16: 252–256

Nicolau B, Netuveli G, Kim JW, Sheiham A, Marcenes W (2007a) A life-course approach to assess psychosocial factors and periodontal disease. J Clin Periodontol 34:844–850

Nicolau B, Thomson WM, Steele JG, Allison PJ (2007b). Life-course epidemiology: concepts and theoretical models and its relevance to chronic oral conditions. Community Dent Oral Epidemiol 35:241–249

Nyyssonen V, Paunio I, Rayala M, Vehkalahti M (1984) Dental caries in the adult population in Finland: I. Prevalence of dental caries. Int J Epidemiol 13:486–490

Oliver RC, Brown LJ (1993) Periodontal diseases and tooth loss. Periodontology 2000 2:117–127

Osborn J, Stoltenberg J, Huso B, Aeppli D, Pihlstrom B (1990) Comparison of measurement variability using a standard and constant force periodontal probe. J Periodontol 61:497–503

Parisotto TM, Steiner-Oliveira C, Silva CM, Rodrigues LK, Nobre-dos-Santos M (2010) Early childhood caries and mutans streptococci: a systematic review. Oral Health Prev Dent 8:59–70

Perara I, Ekanayake L (2008) Social gradient in dental caries among adolescents in Sri Lanka. Caries Res 42:105–111

Peruzzo DC, Benatti BB, Ambrosano GM, Nogueira-Filho GR, Sallum EA, Casati MZ, Nociti FH Jr (2007) A systematic review of stress and psychological factors as possible risk factors for periodontal disease. J Periodontol 78:1491–1504

Petersen PE (2003a) The World Oral Health Report 2003. Oral health programme. WHO, Geneva

Petersen PE (2003b) The World Oral Health Report 2003: continuous improvement of oral health in the 21st century – the approach of the WHO Global Oral Health Programme. Community Dent Oral Epidemiol 31(s1):3–24

Petersen PE, Yamamoto T (2005) Improving the oral health of older people: the approach of the WHO Global Oral Health Programme. Community Dent Oral Epidemiol 33:81–92

Phelan JA, Mulligan R, Nelson E, Brunelle J, Alves MEAF, Navazesh M, Greenspan D (2004) Dental caries in HIV-seropositive Women. J Dent Res 83:869–873

Philstrom BL, Michalowicz BS, Johnson NW (2005) Periodontal diseases. Lancet 366:1809–1820

Pitts NB (1997) Diagnostic tools and measurements – impact on appropriate care. Community Dent Oral Epidemiol 25:24–35

Pitts NB (2001) Clinical diagnosis of dental caries: a European perspective. J Dent Educ 65: 972–978

Pitts NB (2004) ICDAS – an international system for caries detection and assessment being developed to facilitate caries epidemiology, research and appropriate clinical management. Community Dent Health 21:193–198

Polson AM, Caton JG (1985) Current status of bleeding in the diagnosis of periodontal diseases. J Periodontol 56(s):1–3

Reich E (2001) Trends in caries and periodontal health epidemiology in Europe. Int Dent J 51: 392–398

Ronderos M, Pihlstrom BL, Hodges JS (2001) Periodontal disease among indigenous people in the Amazon rain forest. J Clin Periodontol 28:995–1003

Rugg-Gunn AJ, Edgar WM (1984) Sugar and dental caries; a review of the evidence. Community Dent Health 1:85–92

Ryan ME, Camu O, Kamer A (2003) The influence of diabetes on the periodontal tissues. J Am Dent Assoc 134:34S-40S

Scannapieco FA, Bush RB, Paju S (2003a) Associations between periodontal disease and risk for atherosclerosis, cardiovascular disease, and stroke. A systematic review. Ann Periodontol 8:38–53

Scannapieco FA, Bush RB, Paju S (2003b) Associations between periodontal disease and risk for nosocomial bacterial pneumonia and chronic obstructive pulmonary disease. A systematic review. Ann Periodontol 8:54–69

Scheutz F, Matee MI, Andsager L, Holm AM, Moshi J, Kagoma C, Mpemba N (1997) Is there an association between periodontal condition and HIV infection? J Clin Periodontol 24:580–587

Schou L, Wight C (1994) Mothers' educational level, dental health behaviours and response to a dental health campaign in relation to their 5 year old children's caries experience. Health Bull 52:232–239

Scottish Dental Clinical Effectiveness Programme (2010) Prevention and management of dental caries in children. SDCEP, Dundee. http://www.sdcep.org.uk/index.aspx?o=2332. Accessed Oct 2011

Scottish Government (2006) Scottish Index of Multiple Deprivation 2006. Scottish Government, Edinburgh

Scottish Government (2009) The Scottish Health Survey 2008. Scottish Government, Edinburgh

Scottish Intercollegiate Guidelines Network (2000) Preventing dental caries in children at high caries risk, targeted prevention of dental caries in the permanent teeth of 6–16 year olds presenting for dental care. SIGN Publications 47, Edinburgh. http://www.sign.ac.uk/pdf/sign47.pdf. Accessed Oct 2011

Scottish Intercollegiate Guideline Network (2005) Prevention and management of dental decay in the pre-school child – a national clinical guideline. SIGN 83 NHS Quality Improvement Scotland, Edinburgh. http://www.sign.ac.uk/pdf/sign83.pdf. Accessed Oct 2011

Selwitz RH, Ismail AL, Pitts NB (2007) Dental caries. Lancet 369:51–59

Shaw D, Macpherson L, Conway D (2009) Tackling socially determined dental inequalities: ethical aspects of Childsmile, the national child oral health demonstration programme in Scotland. Bioethics 23:131–139

Sheiham A, Nicolau B (2005) Evaluation of social and psychological factors in periodontal disease. Periodontology 2000 39:118–131

Sheiham A, Watt RG (2000) The common risk factor approach: a rational basis for promoting oral health. Community Dent Oral Epidemiol 28:399–406

Shiau HJ, Reynolds MA (2010) Sex differences in destructive periodontal disease: a systematic review. J Periodontol 81:1379–1389

Shuler CF (2001) Inherited risks for susceptibility to dental caries. J Dent Educ 65:1038–1045

Skeie MS, Raadal M, Strand GV, Espelid I (2006) The relationship between caries in the primary dentition at 5 years of age and permanent dentition at 10 years of age – a longitudinal study. Int J Paediatr Dent 16:152–160

Sprod A, Anderson R, Treasure E (1996) Effective oral health promotion. Health Promotion Wales, Cardiff

Stabholz A, Soskolne WA, Shapira L (2010) Genetic and environmental risk factors for chronic periodontitis and aggressive periodontitis. Periodontology 2000 53:138–153

Steele J, O'Sullivan I (2011) Executive summary: Adult Dental Health Survey 2009. The Health and Social Care Information Centre, Leeds. http://www.ic.nhs.uk/statistics-and-data-collections/primary-care/dentistry/adult-dental-health-survey-2009--summary-report-and-thematic-series. Accessed Oct 2011

Steele JG, Treasure E, Pitts NB, Morris J, Bradnock G (2000) Total tooth loss in the United Kingdom in 1998 and implications for the future. BDJ 189:598–603

Stookey GK, González-Cabezas C (2001) Emerging methods of caries diagnosis. J Dent Educ 65:1001–1006

Sullivan HR, Harris R (1958) The biology of the children of Hopewood House, Bowral, NSW. II. Observations extending over five years (1952–1956). 2. Observations and oral conditions. Aust Dent J 3:311–317

Susin C, Kingman A, Albandar JM (2005) Effect of partial recording protocols on estimates of prevalence of periodontal disease. J Periodontol 76:262–267

Sweeney PC, Nugent Z, Pitts NB (1999) Deprivation and dental caries status of 5-year-old children in Scotland. Community Dent Oral Epidemiol 27:152–158

Taylor L (1975) Poverty, wealth, and health, or getting the dosage right. BMJ 4:207–211

Tezal M, Grossi SG, Ho AW, Genco RJ (2004) Alcohol consumption and periodontal disease. The Third National Health and Nutrition Examination Survey. J Clin Periodontol 31:484–488

Thomson WM, Poulton R, Milne BJ, Caspi A, Broughton JR, Ayres KMS (2004) Socioeconomic inequalities in oral health in childhood and adulthood in a birth cohort. Community Dent Oral Epidemiol 32:345–353

Toverud G (1957) The influence of war and post-war conditions on the teeth of Norwegian school children. III. Discussions of food supply and dental condition in Norway and other European countries. Milbank Meml Fund Q 35:373–459

TNS Opinion & Social (2010) Eurobarometer 72.3. Oral health report undertaken at the request of the Directorate General Health and Consumers, European Commission. TNS Opinion & Social, Brussels. http://ec.europa.eu/public_opinion/archives/ebs/ebs_330_en.pdf. Accessed Oct 2011

Treadwell HM, Formicola AJ (2005) Improving the oral health of prisoners to improve overall health and well-being. Am J Public Health 95:1677–1678

Turner S, Brewster L, Kidd J, Gnich W, Ball GE, Milburn K, Pitts NB, Goold S, Conway DI, Macpherson LM (2010) Childsmile: the national child oral health improvement programme in Scotland. Part 2: Monitoring and delivery. Br Dent J 209:79–83

Valaitis R, Hesch R, Passarelli C, Sheehan D, Sinton J (2000) A systematic review of the relationship between breastfeeding and early childhood caries. Can J Public Health 91: 411–417

van der Putten GJ, Vanobbergen J, De Visschere L, Schols J, de Baat C (2009) Association of some specific nutrient deficiencies with periodontal disease in elderly people: a systematic literature review. Nutrition 25:717–722

Vartanian LR, Schwartz MB, Brownell KD (2007). Effects of soft drink consumption on nutrition and health: a systematic review and meta-analysis. Am J Public Health 97:667–675

Verrips GH, Kalsbeek H, Eijkman MAJ (2006) Ethnicity and maternal education as risk indicators for dental caries, and the role of dental behavior. Community Dent Oral Epidemiol 21:209–214

Walsh T, Worthington HV, Glenny AM, Appelbe P, Marinho VCC, Shi X (2010) Fluoride toothpastes of different concentrations for preventing dental caries in children and adolescents. Cochrane Database Syst Rev 1:CD007868

Watt RG, Marinho VC (2005) Does oral health promotion improve oral hygiene and gingival health? Periodontology 2000 37:35–47

Whelton H, O'Mullane DM (2007) Public health aspects of oral diseases and disorders. Dental caries. In: Pine C, Harris R (eds) Community oral health. Quintessence, London, pp 165–176

WHO (1986) The Ottawa Charter for Health Promotion. WHO, Geneva

WHO (2010a) Global oral health database and the WHO oral health country/area profile programme. WHO, Malmo

WHO (2010b) World health organization: significant caries index. http://www.whocollab.od.mah.se/sicdata.html. Accessed Oct 2011

Wong MCM, Glenny A-M, Tsang BWK, Lo ECM, Worthington HV, Marinho VCC (2010) Topical fluoride as a cause of dental fluorosis in children. Cochrane Database Syst Rev 1:CD007693

Wyatt CCL, MacEntee MI (2008) Dental caries in chronically disabled elders. Special Care Dent 17:196–202

Xiong X, Buekens P, Fraser WD, Beck J, Offenbacher S (2006) Periodontal disease and adverse pregnancy outcomes: a systematic review. BJOG 113:135–143

Epidemiology of Digestive Diseases

58

Antje Timmer and Lloyd R. Sutherland

Contents

A. Timmer (✉)
Department of Clinical Epidemiology, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

L.R. Sutherland
University of Calgary, Calgary, AB, Canada

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_49,

58.1 Introduction

The most obvious characteristic of digestive disease epidemiology is diversity. Most other clinical specialties deal with one or a limited number of organs each, and within these specialties, certain diseases feature much more prominently than others. In contrast, the digestive system comprises all parts of the alimentary canal from the esophagus to the anus, several solid organs (liver and pancreas), the biliary tree, and the intra-abdominal fascial structures (mesentery, peritoneum) (Fig. 58.1). As the organs are manifold, so are the ways in which they may be affected by disease – structural, inflammatory, cancerous, functional, psychosomatic, endocrine, toxic/drug induced – the list is non-exhaustive. None of the digestive diseases are so prevalent as to dominate the specialty. Rather, the most common 10 diseases taken together represent only 20% of the total prevalence of the digestive diseases (Everhart and Ruhl 2009a). The remaining 80% are spread out over a large number of less frequent diseases.

Similarly, in digestive disease epidemiology, most epidemiological methods may be applicable. Infection and vaccination, genetic epidemiology, cancer and screening, public health and prevention, nutrition, environment, clinical epidemiology including diagnosis, prognosis and therapy, health economics, pharmacoepidemiology all are relevant issues in gastroenterology. In short, digestive disease epidemiology offers a particularly wide scope of diseases and methods.

Conferences and textbooks of epidemiology rarely feature explicit sections on digestive disease epidemiology. More often, selected topics emerge in other contexts, such as risk factors for colorectal cancer in cancer epidemiology, *H. pylori* transmission in infectious disease epidemiology, and diarrhea prevention

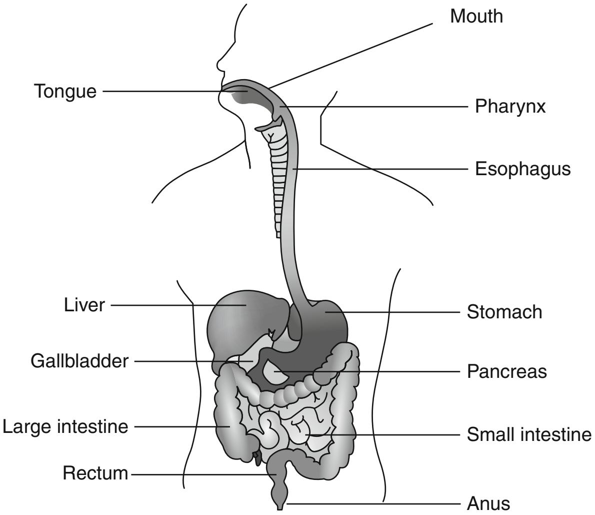


Fig. 58.1 Organs of the digestive system. Courtesy of David Darling, http://www.daviddarling.info/encyclopedia/D/digestive_system.html

and control in developing countries epidemiology. A substantial body of the digestive disease epidemiology literature is generated by clinical or bench-based researchers who have never attended an epidemiology course or conference. Putting this together in a comprehensive way is beyond the scope of a single chapter. Rather, we will highlight a few topics we feel are interesting or of prominent public health relevance. Others are chosen because they hide pitfalls which may evade the epidemiologist not familiar with the clinical aspects of these diseases. For a more comprehensive approach, we refer the interested reader to a book on gastrointestinal epidemiology (Talley et al. 2007).

Following an introduction into the scope and approaches used in digestive disease epidemiology, we will first discuss the relative frequency and burden of digestive diseases, including trends over time (Sect. 58.3). We will move from a global perspective to the descriptive epidemiology of several specific problems representing the main organ groups within the digestive tract. Care has been taken to focus on exemplary, common problems, such as viral Hepatitis and related liver disease, the more common gastrointestinal cancer sites, the inflamed bowel, and functional diseases.

In the fourth section, etiological aspects will be described, covering genetics, poverty-related factors, and behavioral risk factors. *H. pylori* infection will be discussed, as this is maybe the most important discovery in gastroenterology in the last century.

A special section will be devoted to clinical epidemiology, as this field features particularly strong in digestive disease epidemiology. All sections will be illustrated by a number of examples. We will start with an overview on the descriptive epidemiology of digestive diseases.

58.2 Scope and Approaches

Epidemiology of digestive diseases deals with the frequency, distribution, risk factors, clinical behavior, and management of diseases of the digestive tract. Internationally, concepts of what belongs to the corresponding clinical specialty of gastroenterology vary. The main categories as listed in ICD 10 (World Health Organization 2007; http://www.who.int/classifications/icd/en/) are shown in Table 58.1. For this chapter, we will include aspects of the epidemiology of the liver (hepatology) along with diseases of the upper gastrointestinal tract, bowel, and pancreas but exclude metabolic disease, in particular diabetes, and disorders of the oral cavity.

As there is no tradition of digestive disease epidemiology which might have driven specific methods, we will focus on the descriptive epidemiology of several of the more common disorders, risk factor associations, and some observations on the success of preventive measures. For more detailed methodological aspects, we refer the readers to other chapters of this handbook.

Table 58.1 Diseases of the digestive system

ICD 10	Disease
A00–A09	Intestinal infectious diseases
B15–B19	Viral Hepatitis
C15–C26	Malignant neoplasms of the digestive organs
D00, D01, D12, D13, D27	In situ neoplasms, benign neoplasms, unknown behavior of oral cavity and digestive organs
C15–C26	Malignant neoplasms of the digestive organs
K20–K31	Diseases of esophagus, stomach, and duodenum
K35–K38	Diseases of appendix
K40–K46	Hernia
K50–K52	Non-infective enteritis and colitis
K55–K63	Other diseases of intestines
K65–K67	Diseases of peritoneum
K70–K77	Diseases of liver
K80–K87	Disorders of gallbladder, biliary tract, and pancreas
K90–K93	Other diseases of the digestive system

58.3 Global Burden of Digestive Diseases

58.3.1 Overview

The burden of digestive diseases varies worldwide. In the developing world, infectious diseases, including diarrhea and viral Hepatitis, as well as cancer of the liver, are major problems. In the West, functional diseases are more prevalent. With the exception of infectious diseases and cancers of the digestive tract, there are few valid data on population prevalences. As many diseases are non-fatal, ranks based on mortality rates give another picture than figures derived from hospital discharge diagnoses, and the distribution in ambulatory care is again quite different, as are the health economical consequences (Table 58.2).

About one in three citizens attends a physician for digestive complaints each year in the Western world (Everhart and Ruhl 2009b). Gastrointestinal disease is the third most common cause of death, the leading cause of cancer death, and the most common cause of hospital admission in the UK (Williams et al. 2007).

The major causes of death from gastrointestinal disease in high-income countries, excluding cancer, are liver cirrhosis, peptic ulcer, ischemic bowel disease, diverticular disease, and gastrointestinal hemorrhage. Many other, non-fatal diseases of the digestive tract rank relatively low in vital statistics and hospital admission rates but impact substantially on the quality of life of those affected. Typical examples include gastroesophageal reflux disease, dyspepsia, irritable bowel syndrome, and

Table 58.2 Top ten prevalent digestive diseases (USA 2004; Everhart and Ruhl 2009b)

	Ambulatory care	Hospital discharges	Deaths	Total costs
1	Gastroesophageal reflux disease	Gastroesophageal reflux disease	Colorectal cancer	Chronic liver disease/cirrhosis
2	Chronic constipation	Diverticular disease	Liver disease/cirrhosis	Gastroesophageal reflux disease
3	Abdominal wall hernia	Chronic liver disease/cirrhosis	Pancreatic cancer	Colorectal cancer
4	Hemorrhoids	Chronic constipation	Esophageal cancer	Gallstones
5	Diverticular disease	Gallstones	Gastric cancer	Abdominal wall hernia
6	Irritable bowel syndrome	Peptic ulcer disease	Primary liver cancer	Pancreatic cancer
7	Hepatitis C	Pancreatitis	Bile duct cancer	Diverticular disease
8	Colorectal cancer	Gastrointestinal infections	Hepatitis C	Pancreatitis
9	Chronic liver disease/cirrhosis	Hepatitis C	Gastrointestinal infections	Peptic ulcer disease
10	Gastrointestinal infections	Abdominal wall hernia	Peptic ulcer disease	Hepatitis C

anorectal disorders (e.g., hemorrhoids or fecal incontinence). Symptoms of gastrointestinal disease tend to be unspecific, such as abdominal discomfort, loss of appetite, bloating, heartburn, fatigue, or constipation. Many patients never present to a physician, and in many instances, gastrointestinal conditions remain untreated or in patient self-management. There is an increasing body of research on the determinants of health-care-seeking behavior, in particular in functional diseases.

For most diseases, there is no single test or key symptom to prove the diagnosis. This may pose problems in studying these diseases on a population level. Use of administrative data can also be a challenge as many gastrointestinal conditions occur in the presence of other morbidity. As an example, elderly patients on multiple drugs are of particularly high risk to suffer from gastrointestinal hemorrhage, for example, as a consequence of gastric ulcer related to the use of non-steroidal anti-inflammatory drugs (NSAID). Obviously, the accuracy of, for example, hospital discharge or mortality data depends on the number of diagnoses available for analysis per case.

58.3.2 Examples

58.3.2.1 Infectious Diseases: Hepatitis C Still on the Rise

Methodological approaches to infectious disease epidemiology are covered in chapter ▶Infectious Disease Epidemiology of this handbook. Three infectious entities seem particularly important and interesting from an epidemiological and public health view. Besides acute infectious diarrhea and *H. pylori*, which will be discussed below, these are viral hepatic infections, specifically chronic Hepatitis C.

Viral Hepatitis may be caused by a number of viruses, most prominently Hepatitis A, B, or C (HAV, HBV, and HCV). The main differences of these are described in Table 58.3. In summary, HAV infection is a food-borne and self-limiting acute infection. HBV and HCV are blood-borne and may run a chronic course. It is estimated that about one third of the world's population has been exposed to either HBV or HCV, although only about a third of those affected are aware of the infection (Lancet Editorial 2008).

HAV and HBV are, as so many infectious diseases, associated with poverty-related risks, namely poor sanitation (HAV, HBV), unsafe sex, and vertical transmission from infected mothers during delivery (HBV). Both HAV and HBV are endemic in certain low- and middle-income countries and currently on the decrease thanks to successful interventions, including vaccination programs and screening of pregnant women. In high-income countries, HAV and HBV are now mainly restricted to specific risk groups: HAV in international travelers and HBV in immigrants, inmates, and i.v. drug abusers.

Quite in contrast, the burden from HCV is still expected to rise. In most high-income countries, HCV is now the most common cause of chronic liver disease, in others, second only to alcohol. In the USA, the current prevalence of chronic HCV is 1.3% (RNA positive), as compared to 0.4% for chronic HBV (Liangpunsakul and

Table 58.3 Viral Hepatitis – differences A, B, C

Characteristics	HAV	HBV	HCV
Main mode of transmission	Feco-oral: Contaminated food and water Close personal contact	Blood-borne, mucosal: perinatal Sexual intercourse/unsafe sex i.v. drug abuse, needle sharing	Blood-borne: i.v. drug abuse, needle sharing blood products before 1992 Other iatrogenic (contaminated needles)
Other ways of transmission	Homosexual activity i.v. drug abuse	Hemodialysis Blood products in unsafe health care settings Household sharing	Hemodialysis Professional (needle stick injury) Controversial: sexual intercourse (multiple partners) Rare: tattooing, acupuncture, piercing, household sharing, perinatal
Diagnosis	Anti-HAV IgM/IgG – acute or past exposure; Immunity	HBs Ag = acute or chronic infection Anti-HBs = past exposure; immunity	Anti-HCV = past exposure or chronic infection HCV-RNA = confirmatory test (acute or chronic infection)
Vaccination	Available since 1995	Recommended since 1992	Not available
Acute infection	Mostly symptomatic in adults, often no symptoms in children	Mostly symptomatic (jaundice), Rare: fulminant (liver failure)	Rarely symptomatic
Proportion resulting in chronic infection	<1%	<10% (more if perinatally acquired)	55–90%, depending on age and sex
High-prevalence regions	Developing world (>90% exposed during childhood); occasional outbreaks in middle-/high-income countries	Far East, Southeast Asia (up to 16% chronic HBV), sub-Saharan Africa, Amazon area, Eastern Europe	More evenly distributed including North America and Europe (world wide 2.2%)
Number of persons currently infected (worldwide)	Incidence 1.4 million/year (acute)	Prevalence >350 million (chronic)	Prevalence 170 million (estimated) (chronic)
Preventive strategies	Vaccination (children, international travel, high-risk persons) Improve sanitation	Vaccinations (infants and high-risk persons) Screening of pregnant women Immunoprophylaxis to avoid perinatal transmission Screening of donors/inactivation of blood products Safer sex	Screening and testing of blood donors Risk reduction counseling of persons at risk, e.g., avoid needle sharing, safer sex

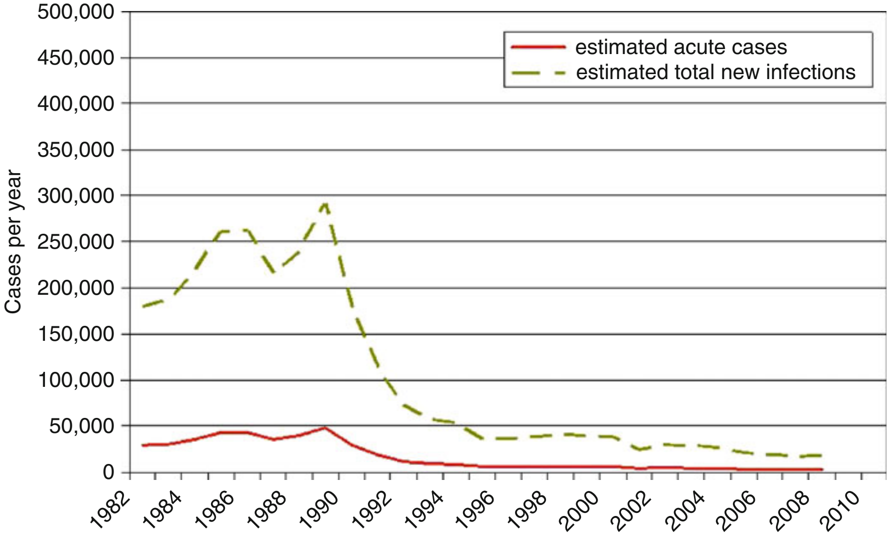


Fig. 58.2 HCV infection – number of cases per year (USA, CDC data). Source: http://www.cdc.gov/hepatitis/Statistics/index.htm (with kind permission)

Chalasani 2005). In Western Europe, HCV accounts for 70% of all cases of chronic Hepatitis, 40% of all liver cirrhosis, and 60% of hepatocellular carcinoma (Williams et al. 2007).

Acute HCV infection is rarely symptomatic and mostly goes unnoticed (World Health Organization 2007). Diagnosis is usually delayed until routine lab examinations unveil raised liver enzymes or until complications of liver disease manifest. In consequence, there are hardly any empirical data on the incidence of HCV. Estimations have been based on mathematical modeling (Davis et al. 2010). These indicate a steep fall in the incidence from the first description of the virus in 1989 to recent times. The decrease has been linked to the introduction of blood donor testing in 1989, improved diagnostic tests in 1992, and the promotion of safer needle using techniques in i.v. drug users (Fig. 58.2).

For the natural history of the disease, several natural experiments are available. In Germany, several thousand women were infected with HCV when immunoglobulin from donors with acute non-A and non-B Hepatitis (as this viral Hepatitis was known before the description of HCV) was used for rhesus incompatibility prophylaxis. Two thousand four hundred and sixty four affected women were acknowledged as vaccination victims and received compensation from the Eastern German Government from 1979 and later from the Federal Republic of Germany. In Ireland, a regional analysis of blood donors in 1991 discovered a strong association of chronic Hepatitis C with rhesus-negativity. In a subsequent national investigation, more than 62,000 of women were tested who had received anti-D immune globulin between 1970 and 1994, confirming that quantities used in 1977 and 1978 had been contaminated. All infections could be traced back to a single donor, and 1,042 successful claims were made by infected women (Akehurst 1999). More diverse claimant cohorts are available for follow-up from other countries, such as the UK, Canada, and Japan (Harris et al. 2002; Pokorski 2001; Thein et al. 2009). Male sex, older age at infection, use of alcohol, and coinfection with HIV or HBV are the most important risk factors for HCV-related morbidity (Poynard et al. 1997).

A prognostic study using complex Markov modeling took into account the different prognosis by age and sex and arrived at the following estimations (Davis et al. 2010): Following acute infection in women under the age of 30, 55% will run a chronic course. Of these 4.2% will have end-stage liver disease (cirrhosis) after 30 years, 0.02% will have hepatocellular carcinoma (HCC), and 0.9% will have died from liver-related death. In contrast, infected male persons aged 31–50 have a risk of 80% for chronic HCV and, if chronic, will suffer from cirrhosis in 38.1%, HCC in 0.5%, and liver death in 11.1%.

Of those currently infected, most contracted the virus between 1970 and 1990 and are now entering the stages of complicated liver disease. According to the Markov model, the prevalence of cirrhosis within those currently infected will continue to increase from currently 25 to 45% in 2030, before, eventually, the burden will decrease. The most important message is: despite the major advances in reducing the incidence of this infection since it was first described, the peak of the HCV epidemic is still to come.

58.3.2.2 Cancer of the Gastrointestinal Tract: Subsite Matters

Cancer epidemiology is covered by chapter ▸Cancer Epidemiology of this handbook. A few aspects specific to gastrointestinal cancers seem worth noting in the context of digestive disease epidemiology. Digestive cancers account for about 18% of all cancers in the Western world. Most prominently, these are the primary malignancies of the esophagus, stomach, large bowel, liver, and pancreas. Less commonly, the biliary tract, the small bowel, or the peritoneal cavity is affected.

For most entities, men are affected more often than women. The average age at onset is relatively high as compared to cancers at other sites. Most malignancies in the gastrointestinal tract are solid, and of these, most are adenomas, although other entities, for example, lymphoma or neuroendocrine tumors, do occur. The majority of gastrointestinal cancers still have a very poor prognosis. For example, less than 5% of those diagnosed with cancer of the pancreas survive 5 years following the diagnosis; mortality rates still equal incidence. The decrease of both incidence and mortality in colorectal cancer (see chapter ▸Cancer Epidemiology of this handbook) is one of the rare successes in gastrointestinal cancer epidemiology and so far mainly restricted to Western countries, specifically those where screening takes place.

Quite in contrast to colorectal cancer, primary cancer of the liver (hepatocellular carcinoma, HCC) is endemic in Africa, where prevalence rates may be as high as 20% (Hainaut and Boyle 2008). This is a direct consequence of a high prevalence of chronic HBV. HCC almost exclusively develops in the presence of cirrhosis of the liver, and in low-income countries, chronic HBV infection is the most common cause of cirrhosis of the liver. Vertical transmission of HBV to newborns of infected mothers carries the highest risk of both running a chronic course of HBV infection as well as of resulting in cancer. HBV is amenable to preventive measures, most importantly childhood vaccination programs, screening of pregnant women, and perinatal active-passive immunization of babies of infected mothers. Consequently, HCC may be expected to decrease in these countries, while it is currently on the rise in the Western world due to the epidemic of HCV-associated morbidity, as discussed above.

Divergent trends in incidence are observed in esophageal cancer. So far, the dominant histological type of cancer of the esophagus has been squamous, but adenomatous cancer is on the rise and is now more frequent in several Western European countries (Bosetti et al. 2008). Squamous carcinoma of the esophagus is closely associated with smoking and alcohol (see below). These risk factors are estimated to account for more than 80% of this highly fatal malignancy (Danaei et al. 2005). In contrast, adenocarcinoma of the esophagus, arising from the lower part of the esophagus, is related to acid reflux and obesity.

Cancer of the upper (cardiac) part of the stomach and the gastroesophageal junction is in its risk profile similar to lower esophageal cancer. The distinction may sometimes be difficult to make. Other than cancer of the lower (non-cardiac) part of the stomach, infection with *H. pylori* does not play a role in cancer of the cardiac region or may even be protective. Whereas the elimination of *H. pylori* will result in a decreased risk for non-cardiac gastric cancer, this is not the case for cancer of the cardiac region. There is even some controversy whether *H. pylori* might be protective for this type of cancer.

This may appear confusing. However, these examples underline the importance of thoroughly defining subtypes of a given cancer in terms of site and histology in order not to blur any risk associations or time trends.

58.3.2.3 The Inflamed Bowel: Inflammatory Bowel Disease, Appendicitis, and Diverticulitis

About every part of the digestive system may be affected by chronic inflammation and autoimmune reactions. Examples include autoimmune pancreatitis, pernicious anemia (arising from autoimmune gastritis), autoimmune Hepatitis, primary biliary cirrhosis, and celiac disease. With the exception of celiac disease (gluten-sensitive enteropathy, 1% of the population), these conditions are relatively rare, and few data are available from larger cohorts regarding risk factors and incidence. For some conditions, viral triggers are suspected. The epidemiological literature is scarce.

In contrast, innumerable descriptive and case-control studies have been published on the incidence and possible etiological factors of the inflammatory bowel diseases (IBD), Crohn's disease, and ulcerative colitis. Typically, these diseases manifest in early adulthood and run a chronic, relapsing course. Complications may include growth failure in the young, bowel obstruction, and fistulization in Crohn's disease frequently necessitating surgery and colorectal cancer in ulcerative colitis. Although life expectancy is now close to normal, morbidity is high. In consequence, these patients constitute a relevant fraction in the daily practice of a gastroenterologist. Inflammatory bowel diseases are also an attractive target for biological therapy, as well as genetic studies, and feature highly in the clinical trial literature. The prevalence in the population is relatively low and is currently estimated to be about 0.4% in North America and Western Europe (Loftus Jr. 2004). These diseases have been linked to a Western lifestyle. They were very uncommon before World War II and have increased since.

Quite in contrast, other inflammatory bowel disorders are more common but receive less research attention. In particular, these are acute appendicitis, which may affect up to 10% of persons in the Western world during a lifetime, mostly at young age, and diverticulitis, a common condition in the elderly. While IBD is a disease of the second half of the twentieth century, appendicitis incidence peaked in the first half of the last century and has been declining since then. Not only do these diseases show divergent time trends in correlational studies, there is also a strong negative association between ulcerative colitis and appendicitis in individual level analyses: the odds ratio (OR) of getting ulcerative colitis in appendectomized persons is about one third of those not appendectomized (Frisch et al. 2009). This finding is so far unexplained. Overall, there has been remarkably little research on appendicitis, although it is such a common surgical condition. Most of the descriptive and etiological literature stems from the 1970s and 1980s (Heaton 1987).

Diverticular disease is another common but relatively understudied condition. Up to 70% of the elderly are estimated to have diverticulosis of the large bowel, little pouches in the colon wall arising from increased intraluminal pressure, associated with a sedentary lifestyle and low-fiber diets. Only a minority of these become symptomatic, but still so many as to make this one of the ten most common digestive diseases (Everhart and Ruhl 2009b). Diverticula of the large bowel are also the most frequent cause of lower gastrointestinal bleeding. In the case of inflammation (diverticulitis), if complicated by abscess, peritonitis, or recurring, the affected part of the bowel will be removed surgically. Thus, as appendicitis, diverticulitis can be cured by surgery. Other than IBD, it is usually not associated with lifelong morbidity.

58.3.2.4 Hepatobiliary and Pancreatic Diseases: Gall Stones, Pancreatitis, and Cirrhosis of the Liver

Gallstones are among the most common gastrointestinal conditions. They are easily diagnosed using ultrasonography, so valid cross-sectional evaluation of representative populations is possible. Ultrasonography-based prevalence data are very variable even within the same country (Kratzer et al. 1999). For example, figures within Germany ranged from 8% to 21% (Völzke et al. 2005; Walcher et al. 2005). This is not yet sufficiently explained.

Risk factors for gallstone disease include age, female sex (parity), obesity, fast weight loss, and some digestive diseases interfering with intestinal resorption, for example, inflammatory bowel disease. There is also a strong genetic component. Gallstones are particularly common in North American natives, where up to two thirds of the female population are affected.

The presence of gallstones does not indicate disease. Most gallstones are asymptomatic and not in need of therapy. It is estimated that only 1% to 2% of persons with gallstones will develop complications per year. Most commonly, these are colics from biliary obstruction or inflammation of the gallbladder (cholecystitis). Usually, treatment is by cholecystectomy.

Gallstones are also the most common cause of acute pancreatitis, as obstruction of the biliary tract may lead to this condition. We will discuss this disease entity in the context of alcohol-related diseases.

Chronic liver disease, in particular cirrhosis of the liver, is among the most expensive disease groups in gastroenterology and among the most deadly (Everhart and Ruhl 2009c). This is the more tragic in that this condition is to a large extent preventable. We have touched on this before: in the developing world, cirrhosis is usually due to chronic HBV infection. In contrast, in high- and middle-income countries, cirrhosis of the liver is most commonly due to HCV infection, followed by alcoholic and non-alcoholic fatty liver disease. There are many other causes including autoimmune and metabolic disorders, but these non-preventable conditions taken together cause less than 20% of all cases of cirrhosis.

About 1% of the population is estimated to have histological evidence of cirrhosis, that is, scarring of the liver resulting from chronic liver injury. Hepatic tissue is replaced by connective tissue as part of the wound healing process; the liver slowly becomes fibrotic. The end stage is characterized, on the one hand, by loss of liver function and on the other hand by increased hepatic resistance resulting in hypertension within the portal vein and circulatory deregulation. Patients with end-stage liver disease may suffer from diverse life-threatening complications, including hepatic coma, bleeding from esophageal varicose veins, peritonitis, and kidney failure. These account for the substantial burden chronic liver diseases pose to society. Mortality is high, and liver transplantation is the only treatment option in advanced liver disease. The pattern of the global burden of liver disease and its causes is expected to change substantially over the coming decades.

58.3.2.5 Functional Diseases: Definitions Matter

Functional diseases of the digestive tract are among the most common disorders in the population, affecting about a third of the population in Western countries (Koloski et al. 2002). Irritable bowel syndrome alone is estimated to constitute up to 50% of the workload of a gastroenterologist working in ambulatory care (Williams et al. 2007). Prevalence rates are mostly unreliable and vary depending on the definitions used (Vandvik et al. 2004). Only a minority of those affected seeks medical care and ever gets a diagnosis, and of those who do see a general practitioner, many are never referred to a specialist.

In general, functional disease is assumed in the presence of symptoms without underlying organic disease. This is overly simplistic. Insufficiently understood gut motility abnormalities seem to arise from disordered interaction between the central nervous system and the gut nervous system (Cremonini and Talley 2004). Learned illness behavior and emotional factors may play a role in how the individual copes with these common symptoms and whether medical help is sought (Levy et al. 2000). Although not usually associated with increased mortality or hospitalization, quality of life is known to be decreased, and substantial loss of work productivity has been shown.

Quite a few functional entities are defined in gastroenterology: globus pharyngis (feeling a lump in the throat), non-cardiac chest pain, functional (or non-ulcer) dyspepsia, functional abdominal pain, functional constipation, and irritable bowel syndrome. Of these, dyspepsia, constipation, and irritable bowel syndrome are the most common. Substantial overlap may exist in the symptoms reported, both between the different functional disorders, as between functional and organic disorders. This makes the definition of diagnostic criteria such a challenge. Dyspepsia, for example, refers to abdominal discomfort, which may or may not be associated with peptic ulcer, gastritis, or esophagitis. In about 50%, there is no structural correlate, and, thus, functional or non-ulcer dyspepsia is assumed (Vakil and Talley 2007). Clinical symptoms are not sufficient to differentiate between these entities (Moayyedi et al. 2006).

Chronic constipation is endemic in certain subpopulations, in particular the elderly. It may constitute a major management problem in nursing homes. In a minority of patients, slow bowel transit or defecatory dysfunction is identified. These patients form a specific subgroup. Constipation is also a side effect of many medications, most notably opioids. Furthermore, it is associated with a variety of neurological and psychiatric conditions such as Parkinson's disease, multiple sclerosis, or depression. The average prevalence in the normal population in North America is 15% if stringent diagnostic criteria are used but may be up to 27% when studies rely on self-reporting (Higgins and Johanson 2004).

Thus, functional disorders pose specific problems for epidemiological research. Not only do diagnosis and symptom severity require careful definitions, the choice of outcome criteria for clinical trials is also a challenge. Most commonly, the Rome criteria are used in research and practice, currently in their 2nd revision (Rome III) (Drossman 2006). The Rome group differentiates 28 adult and 17 pediatric functional gastrointestinal disorders and has developed consensus criteria for those more frequent. Symptom questionnaires are also available, although mostly not validated for use in epidemiological research. The key publications on the diagnostic criteria by disease as well as diagnostic questionnaires and scoring codes are available from the Rome Foundation website (www.romecriteria.org).

58.4 Risk Factors and Determinants, Prevention and Control

58.4.1 Overview

There is a high potential for public health interventions in digestive disease. Hand washing to prevent the spread of acute diarrhea is a prominent example. Many diseases of the alimentary tract are strongly associated with risk behavior, for example, alcohol and i.v. drug abuse in chronic liver disease. Some are amenable to prevention by eradication of precursor conditions, such as adenomas in colorectal cancer or *H. pylori* infection in peptic ulcer disease and gastric cancer. Vaccination programs are also in place, apt to substantially changing the prevalence of several infectious diseases. Examples include Hepatitis B or rotavirus infection.

In this section, we describe some of the major risk factors in gastrointestinal disease by the example of selected entities. Besides poverty in the context of infant mortality, alcohol is probably the most important modifiable risk factor for digestive diseases. It contributes to the majority of cases of esophageal cancer, pancreatitis, and cirrhosis of the liver in affluent societies. A comprehensive analysis of published data calculated population attributable fractions for several common cancers due to nine common environmental and behavioral risk factors (Table 58.4) (Danaei et al. 2005).

Most of the current research in digestive disease etiology seems to focus on genetics.

58.4.2 Examples

58.4.2.1 Genetic Factors: Crohn's Disease as a Prominent Example of Complex Heritability

Several digestive diseases, or multiorgan diseases affecting the digestive system, display a monogenic heritability. Examples include hemochromatosis (affecting the liver, pancreas, and brain), cystic fibrosis (affecting the lung and pancreas), alpha-1-antitrypsin deficiency (affecting the lung and liver), several types of hereditary pancreatitis, hereditary non-polyposis cancer of the colon, and familial adenomatosis coli (FAP). These relatively rare conditions have, in the clinical context, important implications for surveillance programs, as some are precancerous, and others are associated with severe non-malignant, sometimes avoidable morbidity. Counseling of affected families is necessary.

More commonly, complex heritability is present, posing considerable challenges to the effective collaboration between epidemiologists and geneticists. For methodological issues, please refer to chapter ▸Statistical Methods in Genetic Epidemiology of this handbook. Crohn's disease may be taken to illustrate quite a few of the techniques presented there (Gaya et al. 2006; Cho 2008). Observations of increased prevalence in families of affected persons, higher concordance in monozygotics in several twin studies, segregational analysis pointing at an autosomal recessive gene, and mapping of a first susceptibility locus (IBD1) preceded the concurrent identification of the NOD2 gene using positional cloning on the one (Hugot et al. 2001) and candidate gene assessment on the other hand (Ogura et al. 2001) in 2001. This was just a start: as of today (July 2010), a PubMed search using the phrase "inflammatory bowel disease AND genetic" returns 3,268 hits – to go into detail is beyond the scope of this chapter. The many mutations discovered so far explain only a minority of cases.

Epigenetic approaches may help elucidate the pathogenesis not only of Crohn's disease but of gastrointestinal development in general (Waterland 2006). Differentiation of the colonic gut flora taking place in early infancy is suspected by some

Table 58.4 Risk factors for gastrointestinal cancer mortality with population attributable fractions (*PAF*), global (high-/low- and middle-income countries) (Danaei et al. 2005)

	High body mass index (BMI)	Low fruit/vegetable consumption	Physical inactivity	Smoking	Alcohol	Intravenous injection in health-care settings	Risk factors combined
Esophagus (combined)	–	18 (12/19)%	–	42 (71/37)%	26 (41/24)%	–	62 (85/58)%
Stomach (combined)	–	18 (17/19)%	–	13 (25/11)%	–	–	28 (34/27)%
Colorectum	11 (14/9)%	2 (1/2)%	15 (14/15)%	–	–	–	13 (15/11)%
Liver (HCC)	–	–	–	14 (29/11)%	25 (32/23)%	18 (3/21)%	47 (52/45)%
Pancreas	–	–	–	22 (30/15)%	–	–	22 (30/15)%

to play a major role in many diseases, including diabetes and obesity (Frank and Pace 2008; Peterson et al. 2009). Research in this has only just started and may bring the gut much more to the forefront of research, including epidemiology, than previously anticipated.

58.4.2.2 Poor Sanitation, Unsafe Water, and Contaminated Food: Acute Infectious Diarrhea

Several of the digestive diseases are poverty related and occur endemically in countries with crowded housing conditions as well as contaminated food and water supplies. Most of these are infectious (for methods, see chapter ▸Infectious Disease Epidemiology of this handbook). We have already mentioned viral Hepatitis, of which HBV and HAV are amenable to effective preventive measures, including vaccination and screening of pregnant women. *H. pylori* is another example, which we will discuss below. Here we would like to concentrate on an issue of primary importance in global public health: acute infectious diarrhea.

Acute diarrhea is still, besides pneumonia, the leading cause of death in children under the age of five. It accounts for 17% of the infant deaths worldwide (World Health Organization 2007). Diarrhea is common all over the world, in particular in children. Fatality from diarrhea is closely related to the traditional risk factors associated with poverty – the combined risks of underweight, insufficient breastfeeding, and vitamin deficiencies are estimated to account for 73% of infant mortality from diarrhea, and 99% of deaths from diarrhea occur in low-income countries (World Health Organization 2007).

A short summary (not comprehensive) of common causative organisms is presented in Table 58.5 (O'Brian and Halder 2007). Good advances have been made in the prevention and treatment of bacterial and parasitic organisms. Previously common enteric infections such as typhoid fever, cholera, and amoebic colitis are on the decline. Today, most intestinal infections are viral. Of these, norovirus is ubiquitous and regularly causes outbreaks all around the world, affecting persons of any age. In children under the age of five, rotavirus is more common. Besides watery diarrhea, fever and vomiting are also typical, aggravating the risks for the underprivileged. In high-income countries, the disease will subside within a few days. In low-income countries, many children die from dehydration. The relative contribution of rotavirus infection to infant deaths from diarrhea has been increasing and may currently be as high as 40% (Parashar et al. 2009).

The high mortality from acute diarrhea including rotavirus infection is the more tragic as cheap and effective measures are available for treating this condition. WHO oral rehydration solution (ORS) works via the sodium channel glucose transport system in the bowel mucosa. It is superior to i.v. fluids or any other therapy in watery diarrhea (Hartling et al. 2006; Gregorio et al. 2009). Diarrhea prevention programs have been in place from 1978 and have led to dramatic falls in the infant death rate from diarrhea in the following years (Santosham et al. 2010). Unfortunately, in the 1990s, the specific diarrhea programs were merged into integrated health programs. Diarrhea as a distinct entity lost priority, as did the promotion of ORS, formerly named the major medical advance of the twentieth century (Santosham et al. 2010).

Table 58.5 Common risk factors for specific antigens causing infectious diarrhea (Modified from O‘Brian and Halder (2007))

Antigens	Risk factors
E. coli O157:H7	Undercooked beef, unpasteurized milk, apple cider, visits to animal farms, petting zoos
Shigella	Contaminated water and vegetables, day-care centers, custodial institutions
Clostridium difficile	Antibiotics, hospitalized patients
Campylobacter	Undercooked poultry, contaminated milk, tuna salad
Salmonella	Raw eggs, undercooked poultry, unrefrigerated dressing, reptiles as a pet, family members with Salmonella
Non-cholera vibrio	Raw/undercooked seafood
Giardia	Contaminated water, recreational exposure in lakes, rivers, or swimming pools, day-care centers
Cryptosporidium	International travel, contact with cattle, freshwater swimming
Rotavirus	Cold season, infancy
Norovirus	Cold season, raw oyster consumption, cruise ships, health institutions

Since 1995, there has not been much progress in the acceptance of this easy and effective therapy. Currently, only about a third of children with watery diarrhea receive ORS; many health workers still prefer antibiotics for watery diarrhea. If the millennium goal of reducing infant mortality worldwide by two thirds by 2015 is to be met, the prevention and treatment of acute diarrhea need to get back in focus.

A seven-point plan set up for comprehensive diarrhea control includes rotavirus vaccination, promotion of early and exclusive breastfeeding, vitamin A supplementation, promotion of hand washing with soap, improvement of water quantity and quality, promotion of community-wide sanitation, and, for treatment, fluid replacement and zinc supplements (Wardlaw et al. 2010). Two vaccines against rotavirus are now available in developed countries but need to be evaluated in endemic countries as soon as possible. Introduction to low-income countries will depend on priority setting by governments and WHO (Rose et al. 2009).

58.4.2.3 Peptic Ulcer Disease and *H. Pylori*: A Success Story?

One of the most exciting milestones in gastroenterology in the twentieth century was the discovery that peptic ulcer disease was caused by a bacterial agent, *Helicobacter pylori* – rightly awarded the Nobel Prize in Medicine in 2005 (Lang 2005). This major success story of clinical medicine was, quite in contrast, prominently discussed as one famous example where epidemiology failed (Davey-Smith and Ebrahim 2001).

For most of the twentieth century, peptic ulcer disease was a common problem with high morbidity and mortality. Severe complications occurred frequently, such as life-threatening bleeding or perforation into the abdominal cavity. If not fatal, recurrence was almost inevitable. Many patients ended up having major surgery, such as total or, more often, subtotal gastrectomy or vagotomy. These procedures are themselves associated with high perioperative morbidity and subsequent decreased quality of life.

There have been numerous epidemiological studies into the etiology of peptic ulcer disease (Davey-Smith and Ebrahim 2001). This was sparkled, among other things, by so far unparalleled patterns in the incidence and prevalence of these diseases, similar as, more than 50 years later, the inflammatory bowel diseases. Comparable to coronary heart disease, lung cancer, and appendicitis, the incidence of both gastric and duodenal ulcer increased during the first half of the twentieth century. From the 1950s on, however, the incidence of peptic ulcer decreased for unknown reasons. Strong birth cohort effects were discovered as early as 40 years ago, pointing at the relevance of events occurring early in life (Susser 1982; Sonnenberg 2007). Persons born between 1870 and 1920 had the highest risk. For each subsequent generation, the risk for peptic ulcer incidence and mortality successively decreased.

Similar observations for, for example, tuberculosis incidence could have raised the suspicion of an infectious etiology in peptic ulcer disease. It did not. Up until the 1970s, diet, alcohol, emotional stress, and personality had been identified as risk factors for peptic ulcer disease – none of them would have explained the birth cohort effect. Davey-Smith in his critical essay on our discipline points out that the epidemiological transition from infectious to non-infectious diseases may have had, in this case, clouded the view of epidemiologists in favor of non-infectious factors (Davey-Smith and Ebrahim 2001). Improved sanitary conditions and less domestic crowding resulting in lower infection rates would have been a plausible explanation of the birth cohort phenomenon. Still, etiological hypotheses focused on nutrition and psychosomatics.

Epidemiology was not the only discipline led astray in this case. The notion that peptic ulcer disease was caused by microbes colonizing the gastric mucosa was revolutionary in clinical medicine, pathology, and microbiology as well. Until then, the stomach had been assumed to be sterile, due to its high acidity.

It was a junior gastroenterology fellow who performed a simple cross-sectional study in 100 patients referred to endoscopy after a pathologist had described a curved bacterium, later called *Helicobacter pylori*, in the gastric mucosa of patients with gastritis (Marshall and Warren 1984). He found a strong association between the presence of the microbe and active chronic gastritis, duodenal ulcer, or gastric ulcer. Later, he demonstrated acute gastritis in himself after ingestion of *H. pylori*.

H. pylori gastritis is a chronic inflammation as a reaction to *H. pylori* colonization, playing a pivotal role in the development of several gastric disorders. Carriers of *H. pylori* have a 10% to 20% lifetime risk of peptic ulcer disease and a 1% to 2% life time risk of gastric cancer. Taking another perspective, more than 70% of all gastric cancer is attributable to *H. pylori* (Goodwin et al. 1997). Eradication of *H. pylori* with combination of antibiotic therapy effectively cures peptic ulcer disease. It is also expected to decrease the risk for gastric cancer.

Naturally, the question now arises whether screening for and treatment of *H. pylori* are useful from a public health perspective. Several issues play a role in this ongoing discussion – the prevalence of the infection in a population, the availability of tests and their performance, costs, acceptance, and rates of microbial resistance to antibiotic agents.

On average, about 50% of a population carries *H. pylori*. This varies by age and region. In developing countries, prevalence rates may be as high as 80%. In industrial countries, the rate is, on average, 40%, but with a marked generational gradient. Infection seems to occur in the first years of life, primarily from close family contacts. In adults, new infection is very rare, as is spontaneous seroconversion. This means that the observed differences by age group are almost completely caused by generation (birth cohort) effects. Crowding during industrialization and decreasing hygiene has been suggested to explain the effect (Sonnenberg 1995). The prevalence in children is a good indicator for the future burden. German school children now have prevalence rates as low as around 10% (Rothenbacher et al. 1999).

58.4.2.4 Behavioral Risk Factors: Alcohol and Smoking

Four entities in particular are strongly associated with high alcohol intake: esophageal cancer, acute and chronic pancreatitis, and cirrhosis of the liver.

As pointed out above, in esophageal cancer, the type of cancer needs to be defined in order to correctly interpret risk data. For example, in a recent prospective cohort study, the relative risk for squamous esophageal cancer in persons with high alcohol intake ($\geq$30 g/d) was 4.6 (95% confidence interval (CI) 2.2–9.5) as compared to abstainers, but no association was found for adenocarcinoma of the esophagus (Steevens et al. 2010).

Alcohol plays the major role in the pathogenesis of acute pancreatitis, ahead of biliary causes (obstruction by gall stones). The exact mechanism by which alcohol exerts this effect is unclear. Several hypotheses have been discussed, including direct toxic effects of acetaldehyde on acinar cells, sensitization of the pancreas to injury, or mediation via alcohol-induced duodenitis (Chowdhury and Gupta 2006). A genetic predisposition to pancreatic injury from alcohol seems to be essential. Only a minority of heavy drinkers develops pancreatitis.

In acute pancreatitis of any etiology, systemic reactions are common, including multiorgan failure. Case fatality is now about 10%. Acute pancreatitis is also a relevant side effect of some drugs (e.g., occurs in about 3% of those on the immunosuppressive azathioprine) or diagnostic procedures involving the pancreatic duct (endoscopic retrograde cholangiopancreatoscopy (ERCP), occurs in up to 5% of procedures), but these account for only a small part of the overall burden. Other rare causes are viral, traumatic, autoimmune, or hereditary. What makes some people's pancreas susceptible to damage by alcohol is still subject of research. Both genetic and environmental factors seem to play a role (Haber et al. 1995; Yadav and Whitcomb 2010).

Chronic pancreatitis is probably not a consequence of recurrent or intractable acute pancreatitis but due to different yet unclear mechanisms. Up to 80% of cases are estimated to be due to chronic abuse of alcohol. Pancreatic tissue is replaced by scarring; characteristic features are chronic pain and loss of weight, in advanced cases diabetes due to destruction of islet cells (endocrine function) and maldigestion due to loss of exocrine function. Chronic pancreatitis is a major risk factor for pancreatic carcinoma making the latter another mediate, potential consequence of alcoholism.

In the liver, different pathologies may occur following alcoholic injury. Alcoholic fatty liver is the earliest stage, in which lipids are deposited in hepatic cells, and is found in 90% to 100% of liver biopsies in heavy drinkers (Dufour 2007). Alcoholic Hepatitis affects about 10% to 35% of drinkers at some time. It may present very acutely and may result in acute liver failure. Recurrent hepatitic episodes and progression of fatty deposits will result in fibrosis and eventually cirrhosis, in up to 20% of heavy drinkers or 70% of drinkers who have had alcoholic Hepatitis (Mandayam et al. 2004). Alcohol is also an important cofactor for liver damage from other causes, such as chronic viral Hepatitis, where it exerts synergistic effects. Ultimately, cancer may occur in the cirrhotic liver, as in cirrhosis of other etiologies.

As compared to alcohol, smoking is less important in gastrointestinal disease. Several disorders show elevated risks, but for most, the associations are moderate (Steevens et al. 2010). Smoking is, however, an important cofactor of alcohol in several entities – both in esophageal squamous carcinoma as well as in acute pancreatitis, smoking increases the deleterious effect of alcohol.

In inflammatory bowel disease, contrasting roles for smoking have been demonstrated by type of disease. Smoking is the only established external risk factor for Crohn's disease and is, in addition, associated with a poor prognosis (Calkins 1989; Birrenbach and Bocker 2004). In contrast, in ulcerative colitis, smoking seems to be protective. Typically, the disease begins in former smokers, following smoking cessation (Corrao et al. 1998). Attempts to use nicotine (gums, patches) in the treatment of ulcerative colitis have been unsuccessful, partly due to incompliance as side effects are common with nicotine. Although these divergent associations have been known since the 1980s, the underlying mechanisms remain obscure.

58.4.2.5 Risk Factors in Affluent Societies: Sedentary Lifestyle, Nutrition, and a Sheltered Childhood

Colorectal cancer is commonly perceived as a malignant disorder particularly closely associated with the sedentary Western lifestyle. High-calorie food intake low in fruits, vegetables, and fibers and high in processed meats has been confirmed as risk factors (Bingham et al. 2003; Norat et al. 2005; Friedenreich et al. 2006; van Duijnhoven et al. 2009). However, the joint behavioral risk factor attributable fraction is relatively low in colorectal cancer as compared to the other common digestive tract cancers (Table 58.4). It constituted only 13% in a cumulative analysis based on published data and results from the WHO Comparative Risk Assessment Project (http://www.who.int/whr/2002/en/) (Danaei et al. 2005).

In contrast, non-alcoholic fatty liver disease (NAFLD) and non-alcoholic steatohepatitis are increasingly recognized as sequelae of overnutrition in affluent societies. They are now considered the digestive disease manifestation of the metabolic syndrome. Exact mechanisms are still unclear. On biopsy, features resemble those of alcoholic liver disease. Prevalence data depend on the method used for detection (ultrasonography, magnetic resonance imaging (MRI), or liver biopsy). Prevalence has been reported as up to 30% of the US population (Browning et al. 2004). NAFLD may progress and is expected to become the most common cause of cirrhosis and HCC in Western societies once HCV-related complications have peaked and started to regress (Davis et al. 2010). Currently, a lot of research, bench based, as well as clinical, is addressing this relatively newly discovered phenomenon. New diagnostic tools, in particular the elasticity liver scan, may make this disease easier to examine in large-scale population-based studies.

For data on the importance of nutrition on digestive cancers, we refer to chapter ▶Cancer Epidemiology of this handbook and Table 58.4.

In inflammatory bowel disease, a lot of effort has been devoted to the identification of nutritional risk factors, compromised by recall bias and problems of reverse causation. Due to the relative scarcity of the diseases, most investigators relied on retrospective studies. Since long time periods may elapse between the onset of symptoms and first diagnosis, eating patterns are usually already altered once patients become available for interview. There are now a few prospective studies looking into risk factors for IBD as secondary research aims, in particular EPIC (Hart et al. 2008). Unfortunately, as this cohort is conceptionalized for the investigation of causes of cancer, the population is rather old, and only an atypical subgroup of IBD patients presenting late is available.

More interestingly, in analogy to findings for atopic diseases, IBD seem to be associated with a “sheltered childhood,” that is, improved hygienic conditions in infancy (Timmer 2003). Most likely, in genetically predisposed persons, insufficient priming results in an inappropriate immunological response to antigens, probably within the normal gut flora later in life (Baumgart and Carding 2007).

58.5 Clinical Epidemiology of Digestive Diseases

58.5.1 Overview

Clinical epidemiology and, as its application in clinical medicine, evidence-based medicine have a relatively long and successful history in the treatment of digestive diseases. This is reflected by a growing number of clinical practice guidelines taking into account the results of clinical research. Mega-trials, as encountered in cardiovascular diseases, are uncommon, making meta-analyses more important. Quality of life issues are particularly important in functional disorders, and there has been substantial work to develop validated instruments to measure patient-related outcomes of health interventions. However, there are still many underdeveloped areas in clinical epidemiology research.

Diagnostic and therapeutic options available to highly skilled endoscopists and ultrasonographers advance rapidly, and so must be the rigor in assessing the benefit of these procedures. Prognosis and surveillance in high-risk populations are prominent issues, in particular the prevention or early detection of complications in chronic diseases. The clinical strategies to deal with these risks are often insufficiently corroborated by epidemiological data.

Lastly, pharmacoepidemiology in clinical digestive disease epidemiology is very important. Side effects on the digestive tract are the most common side effects of drugs, ranging from slight nausea to life-threatening gastrointestinal bleeding.

58.5.2 Examples

58.5.2.1 Therapy: Randomized Controlled Trials and the Cochrane Collaboration

The evidence base for therapy in digestive diseases varies substantially by entity. There is extensive clinical trial activity for IBD, HCV, and cancer. In contrast, for some of the more common conditions, the evidence base is surprisingly small. As an example, a PubMed search for randomized clinical trials (publication type) renders 1,104 trials for inflammatory bowel disease, but only 55 for diverticulitis (July 2010). For the latter, potential treatment options are relatively cheap. For example, a diet rich in fiber is commonly recommended to prevent recurrence despite absence of evidence of efficacy. In contrast, immunomodulating drugs as applicable in IBD, the duration of interferon therapy in viral Hepatitis, or individualized (antibody) therapy in colorectal cancer have severe commercial as well as health economical consequences.

For the treatment of HCV, combination therapy with interferon and ribavirin is effective (Brok et al. 2010). The virus can now be eliminated in about 50%, depending on viral genotype, age, sex, and coinfections, for example, with HBV or HIV. Very complex treatment recommendations tailor length of treatment to the various clinical situations (Mangia and Andriulli 2010), mostly supported by evidence from clinical trials and observational prognostic studies. Stepping back from the clinical context to a population perspective, the situation is sobering. In the Markow model cited above, treatment of Hepatitis C by combination of antiviral therapy was estimated to reduce the future risk of late complications (cirrhosis, death, HCC) in those infected today by just 1% (World Health Organization 2007). The problem is that an estimated 70% of infections remain undetected and only 25% of all detected cases are offered treatment. The effect of antiviral therapy on the prevalence on HCV-related morbidity will remain small unless awareness both of the infection and of the treatment options is improved along with the advances in efficacy.

Inflammatory bowel diseases are a good example for difficulties encountered when choosing appropriate endpoints for clinical trials in non-fatal diseases. There is considerable interindividual diversity of symptoms and signs which may indicate relapse or a severe disease course. No single phenomenon has been identified which might function as a useful outcome criterion in clinical trials on inflammatory bowel disease, in particular in Crohn's disease. Instead, a number of composite indices have been developed to allow for standard assessment across patients with varying symptoms. The combination of time in or to remission, based on predefined thresholds of well-validated composite clinical activity scores, and the additional assessment of quality of life have been recommended for use in clinical trials (Sandborn et al. 2002; D'Haens et al. 2007). Unfortunately, at this time, there seems to be some falling back on a preference for surrogate measures of (so far) unproven prognostic relevance. They are not even easily assessable nor without risk for the patient. We allude to "mucosal healing," which requires endoscopy for visualization and is poorly correlated with patient symptoms (Rutgeerts et al. 2007).

Clinical trials are even more challenging in the functional diseases. In these disorders, not just the outcome criteria but also diagnosis and inclusion criteria require thoughtful definitions in order to avoid misclassification and advance generalizability.

Of specific concern are the new invasive procedures, such as endoscopic mucosal resection for early cancers of the esophagus and stomach or natural orifice transluminal endoscopic surgery (NOTES) for intra-abdominal operations. These emerging techniques require highly skilled interventional gastroenterologists or abdominal surgeons proficient in endoscopic techniques. Slow learning curves and restriction to highly specialized centers make these techniques difficult to evaluate, even more so than was previously the case for laparoscopic techniques (Sauerland et al. 2004; Keus et al. 2006).

As mega-trials are rare in digestive disease due to the relative rarity of most conditions, meta-analysis strongly features in gastroenterology and hepatology. Inflammatory bowel diseases were the first digestive diseases to be represented by a Cochrane review group. A Cochrane review group is a team within the Cochrane Collaboration which serves as an editorial basis for systematic reviews prepared for a specific group of diseases (see chapter ▶Clinical Epidemiology and Evidence-Based Health Care of this handbook). Currently, there are 52 of these groups worldwide. Four of them are devoted to digestive tract diseases (Table 58.6). Some disorders may be covered by other groups, such as acute diarrhea by the Infectious Diseases Group and fecal incontinence and rectal prolapse by the Incontinence Group. All groups keep specialized registries of clinical trials in the respective area, including hand search activities, screening conference abstracts, and journals not currently indexed in Medline, representing a useful resource for clinical researchers.

58.5.2.2 Diagnosis: Endoscopy and Symptom Scores

Many gastrointestinal diseases depend on invasive procedures for a definite diagnosis, making these diseases difficult to study in epidemiology. For example, in gastroesophageal reflux disease (GERD), there are structural correlates (esophagitis, acid reflux) to define the diagnosis, but these require upper endoscopy and, possibly, pH-metry.

In functional disease, such as non-ulcer dyspepsia and irritable bowel syndrome, there are no diagnostic signs. These are rather syndromes, requiring thoughtful definitions as discussed above. Even where internationally agreed on definitions are

Table 58.6 Cochrane groups covering digestive disease therapy and diagnosis

Name (year founded)	Location, web address	Scope
Inflammatory Bowel Diseases (1995) and Functional Bowel Disorders Group (renamed 2005)	London, Ontario, Canada http://www.cochrane.uottawa.ca/ibd/	Ulcerative colitis Crohn's disease Since 2005: other intestinal disorders including microscopic colitis, collagenous colitis, lymphocytic colitis, irritable bowel syndrome, chronic constipation, and diarrhea
Hepatobiliary Diseases Group (1996)	Copenhagen, Denmark http://ctu.rh.dk/chbg	Hepatic diseases, including alcoholic liver disease, Hepatitis B and C, cirrhosis of the liver, hepatic malignancies Biliary diseases, including gallstone diseases, infections, malignancies
Colorectal Cancer Group (1998)	Copenhagen, Denmark http://cccg.cochrane.org/	Colorectal, anal, and small bowel cancer Benign proctologic diseases Benign peritoneal diseases Surgical treatment for inflammatory bowel diseases Surgical anal diseases Abdominal hernias Appendiceal diseases Diverticulitis Peritonitis
Upper Gastrointestinal and Pancreatic Diseases Group (1998)	Hamilton, Ontario, Canada http://fhs.mcmaster.ca/ugpd/index.html	Disorders of the esophagus, stomach, duodenum, and pancreas

in place such as the Rome III criteria, these have a tendency to undergo frequent revisions and are subject to severe interrater disagreement (Vandvik et al. 2004).

Advanced diagnostic techniques develop rapidly in gastroenterology at this time. More and more new modifications become available, posing new challenges both with respect to the manual skills of the investigators, as to the technical equipment. Examples include contrast-enforced ultrasonography, the endoscopic capsule, virtual colonoscopy, endoscopic ultrasonography, chromo-endoscopy, virtual chromo-endoscopy, magnetic resonance imaging (MRI) of the small bowel or biliary-pancreatic ducts, and so forth. There is some risk that the technical and manual advances are not accompanied by thoughtful evaluation of diagnostic test accuracy, not to speak of the critical assessment of whether the procedure has an effect on outcome. At times, the ultimate goal of diagnosis (or any clinical maneuver) – make the patient better! – gets out of focus when enthused clinicians formulate clinical guidelines. More input from clinical epidemiologists would be desirable to help researchers adhere to good standards as set out in chapter ▶Clinical Epidemiology and Evidence-Based Health Care of this handbook.

58.5.2.3 Prognosis: Early Detection of Long-Term Complications of Gastrointestinal Disease

Gastroenterology and hepatology comprise many conditions associated with severe, potentially avoidable long-term consequences, some of them malignant. Generally in these conditions, the question arises if and how surveillance is appropriate to prevent or alleviate these complications.

Few conditions will inevitably result in cancer if untreated. An example is familial adenomatous polyposis (FAP) – all affected will develop colorectal cancer at an early age, and total proctocolectomy is recommended in the teenage years. More often, the risks are less well defined.

Long-standing extensive ulcerative colitis is associated with an increased risk of colon cancer. Data are conflicting, but cancer rates of up to 30% after 20 years have been reported (Eaden et al. 2001). Most clinical guidelines recommend regular surveillance colonoscopy from about 10 years after the diagnosis, depending on the extent of the inflammation in the bowel. More recent data have given rise to some controversy whether the risk is really that high.

There is a biliary condition associated with ulcerative colitis, primary sclerosing cholangitis, which is itself strongly associated with cholangiocarcinoma, but surveillance is more difficult. Visualization of the biliary system by endoscopic retrograde cholangiography (ERC) or the MRI equivalent (magnetic resonance cholangiography (MRC)) is required, with unproven, maybe improvable, effectiveness.

As pointed out before, cirrhosis of the liver may result in hepatocellular carcinoma (HCC); in fact, very few HCC develop in a non-cirrhotic liver. Tumor markers (alpha-fetoprotein, AFP) and ultrasonography used to be recommended for follow-up. More recently, AFP has been found of insufficient specificity and sensitivity and has been replaced by imaging techniques only (once yearly). The effectiveness of these schemes is still unclear.

Other examples include Barrett's esophagus, which corresponds to displaced gastric mucosa in the low esophagus. From Barrett's metaplasia, adenocarcinoma of the gastroesophageal junction or lower esophagus may develop. Most endoscopists would probably recommend regular follow-up of patients with Barrett's esophagus diagnosed via upper endoscopy, in particular if a longer segment is involved. However, this has, so far, not been shown to be effective. Most cases of Barrett's are undetected, as this condition, as GERD, correlates poorly with symptoms. As Barrett's metaplastic areas can now be removed by endoscopic resection, the role of screening by upper endoscopy might be reconsidered. So far, population screening for upper gastrointestinal disease is not recommended, and individual surveillance is only contemplated when Barrett's mucosa was detected by endoscopy performed for symptoms.

There are several conditions with markedly raised incidence ratios but low absolute risks of lymphoma or other malignancies. These include small bowel malignancy in celiac disease or Crohn's disease. Surveillance is not recommended in these cases. Similarly, screening for pancreatic cancer in chronic pancreatitis (which is a risk factor for pancreatic cancer) or new onset diabetes (which may be a symptom of advanced pancreatic cancer) is not helpful to improve survival in pancreatic cancer. And lastly, there is the unanswered question on how to proceed following a diagnosis of an *H. pylori* infection in asymptomatic patients, as discussed above. So far, most gastroenterology societies do not recommend eradication to prevent gastric cancer, but the issue is still controversial.

In short, many conditions await critical assessment to help decide on the usefulness of surveillance programs. At this moment, for most cases, patients and clinicians are left with considerable uncertainty.

58.5.2.4 Assessment of Quality of Life in Persons with Digestive Diseases

Several of the more prevalent chronic gastrointestinal diseases have mortality rates close to those of the normal population but impact heavily on persons' daily functioning and life satisfaction. Quality of life research aims at assessing this impact, preferably from the patient's view (chapter ▶Health Services Research of this handbook). Health-related quality of life (HRQL) assessment is now firmly established in gastroenterological research (Glise and Wiklund 2002). In functional diseases, the impact of symptoms on HRQL may be the only measure that defines disease status. In chronic inflammatory disease and cancer, HRQL is now routinely used as an adjunct outcome measure in therapeutic trials. A good example is the inflammatory bowel disease questionnaire (IBDQ), developed by Guyatt et al. (1989). HRQL assessment is also important for health economical analyses and pharmacoeconomical evaluations.

Many instruments are available for use in digestive disease epidemiology, some of them well validated and translated to various languages (Borgaonkar and Irvine 2000). Most are multi-item instruments, appreciating aspects of physical well-being and functioning as well as emotional and social consequences of health. Generally, they can be classified as generic or disease specific, symptom or function oriented, and treatment specific (Wiklund 2007).

Generic instruments help to compare quality of life across various diseases or to the general population (Coons et al. 2000). With respect to physical symptoms, generic instruments tend to focus on issues of self-care and mobility – aspects that may be particularly important in joint disease or cancer, but would be expected of limited relevance in most gastrointestinal diseases. Even so, several of these instruments have been shown to work remarkably well in gastroenterology. Most often, the SF-36 and various modifications of this score have been used (Ware Jr. and Sherbourne 1992; Ware Jr. et al. 1996; Bernklev et al. 2005; Billingsley et al. 2007). Also, the 5-item EuroQol has been successfully validated in patients with various bowel and liver diseases (König et al. 2002; Spiegel et al. 2009). This instrument is particularly helpful in that it facilitates the calculation of health utilities (Stark et al. 2010).

Disease-specific quality of life instruments in gastroenterology are numerous. The Patient-Reported Outcome and Quality of Life Instruments Database (ProQolid, http://www.proqolid.org/) currently lists 30 different instruments for digestive diseases excluding cancer. A detailed overview on available instruments has been published a few years ago and is recommended for reference (Borgaonkar and Irvine 2000). The European Organization for Treatment of Cancer (EORTC) promotes a modular approach to assessing quality of life in clinical trials. The core instrument QLQ-C30 assesses quality of life in cancer patients in general and is available in 81 languages (Aaronson et al. 1993). Additional modules have been developed for specific cancer entities, including many gastrointestinal sites (Sprangers et al. 1993). These can be found at the official EORTC website (http://groups.eortc.be/qol/index.htm).

58.5.2.5 Adverse Events, Pharmacoepidemiology, and Health Economics

Some of the most frequent side effects of drugs relate to the digestive system: nausea and vomiting, diarrhea, and constipation. All of these may have grave consequences for the management of patients, in particular in the elderly and in patients with cancer. Here we would like to focus on two other problems: NSAID-induced gastrointestinal bleeding and the various forms of hepatotoxicity.

Following the introduction of successful treatment schemes for *H. pylori*-associated associated peptic ulcer disease, drug related gastric lesions are now the most frequent reason for ulcer bleeding. The risk is particularly high with aspirin, followed by conventional NSAID’s (Hawkey 2009). As often the elderly and the multimorbid are affected, mortality from gastrointestinal hemorrhage is substantial. The prevention of cardiovascular events by antiplatelet therapy on the one hand and the risk of gastrointestinal bleeding on the other hand need thoughtfully balanced consideration. The issue is further complicated by suspected interactions between antiplatelet (clopidogrel) and antiulcer treatment (proton pump inhibitors).

The short history of selective COX-2 inhibitors is an interesting example on how perspectives differ by clinical specialty. These drugs were expected to alleviate the substantial problem of severe gastrointestinal bleeding in patients depending on NSAID therapy. Most of these are patients with joint disease, often elderly. The drugs decreased ulcer bleeding but increased cardiac events.

Critics have noted the paradoxon that

> “COX-2 inhibitors and NSAIDs were found to increase risks of myocardial infarction and liver failure, while NSAIDs additionally caused hundreds of thousands of deaths from ulcer complications worldwide, (but) it were COX-2 inhibitors rather than NSAIDs that fell from favour” (Steinfeld and Bjorke 2002; Hawkey 2009).

As a side note, there are interesting methodological issues relating to these associations. For example, pharmacovigilance studies came to different conclusions as compared to randomized clinical trials with respect to the bleeding potential of selective COX-2 inhibitors due to insufficient control for confounding. This clinical problem has been used to illustrate the usefulness of propensity scores (Schneeweiss et al. 2006).

Drug-related liver toxicity is an interesting field in pharmacoepidemiology and a somewhat complex one. Different mechanisms apply: cell injury, cholestatic patterns or enzyme induction, and hypersensitivity. The clinical relevance may range from asymptomatic, transient liver enzyme elevation to acute life-threatening situations requiring urgent liver transplantation. Most cases are not reported, so the true incidence is unknown. In the USA, hepatotoxicity is the most frequent reason for the Food and Drug Administration to take regulatory action (Navarro and Senior 2006). A recent example is troglitazone used for glycemic control in patients with diabetes, which was withdrawn in 2000 because of an increased risk of acute liver failure.

With respect to costs of medication, proton pump inhibitors are high up the list. Indications include heartburn and GERD, peptic ulcer disease, gastritis, and ulcer protection with antiplatelet therapy. Many patients use this medication lifelong.

There are quite a few situations where cost-oriented decision analyses have the potential to guide treatment recommendations. Besides lifelong drug treatment of GERD as opposed to surgical interventions (fundoplicatio), these concern, for example, screening and treating of *H. pylori*-positive persons in the absence of symptoms or inflammation, the treatment of viral Hepatitis in subgroups of patients with low chances of virus elimination, and the introduction of vaccines. Most of these issues have not been resolved.

58.6 Conclusions

In this chapter, we have highlighted the burden of various digestive diseases and some of the most important risk factors. Some previously important disorders are currently on the decline, some for unknown reasons, some due to improved living conditions, and some due to successful public health interventions. The retreat of peptic ulcer disease is most likely to fall in the category of hygiene-related diseases; the infectious agent has been identified (*H. pylori*). For appendicitis, the etiology is still unclear, and remarkably little research has been published although this is such a common condition. Good examples for successful public health interventions include screening for colorectal cancer and vaccination to prevent Hepatitis B virus infection. Squamous esophageal cancer is also on the decline, due to the decreasing prevalence of smoking and alcohol abuse in Western populations.

For other diseases, the burden is still expected to increase. These are, for example, the long-term sequelae of chronic Hepatitis C infection, non-alcoholic fatty liver disease as a previously undescribed manifestation of the metabolic syndrome, adenocarcinoma of the esophagus and gastroesophageal junction, and, possibly, functional disorders as well as those conditions associated with old age – NSAID-induced gastrointestinal hemorrhage, diverticulitis, and constipation, to name just a few.

A large proportion of the presented problems are preventable. There are strong associations with modifiable risk factors. Specifically, these include alcohol and i.v. drug abuse in the economically advanced societies and infections for which vaccinations and simple sanitary measures are available (Hepatitis B, rotavirus) in the developing world.

Digestive disease epidemiology seems to be difficult to grasp as an entity, as the conditions are so diverse. Also, there are no epidemiological methods specific to gastroenterology. However, we hope that the gastroenterologists' perspective is helpful to bring some issues back into focus that may have run the risk of losing priority. Most urgently, this refers to acute diarrhea in low-income countries.

With the gut flora as the major part of the human microbiom increasingly acknowledged as potentially crucial in the pathogenesis of various disorders, both gastrointestinal as well as others, the digestive system may become more prominent in the study of health in the future. More input from methodologically advanced epidemiology would certainly be of help in many clinical decisions, such as surveillance of clinical populations at high risk for malignancy.

References

Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, Filiberti A, Flechtner H, Fleishman SB, de Haes JC (1993) The European Organization for research and treatment of cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 85:365–376

Akehurst C (1999) Hepatitis C virus infection from contaminated anti-dimmune globulin in Ireland. Eurosurveillance 3:1411

Baumgart DC, Carding SR (2007) Inflammatory bowel disease: cause and immunobiology. Lancet 369:1627–1640

Bernklev T, Jahnsen J, Lygren I, Henriksen M, Vatn M, Moum B (2005) Health-related quality of life in patients with inflammatory bowel disease measured with the short form-36: psychometric assessments and a comparison with general population norms. Inflamm Bowel Dis 11:909–918

Billingsley KG, Morris AM, Dominitz JA, Matthews B, Dobie S, Barlow W, Wright GE, Baldwin LM (2007) Surgeon and hospital characteristics as predictors of major adverse outcomes following colon cancer surgery: understanding the volume-outcome relationship. Arch Surg 142:23–31

Bingham SA, Day NE, Luben R, Ferrari P, Slimani N, Norat T, Clavel-Chapelon F, Kesse E, Nieters A, Boeing H, Tjonneland A, Overvad K, Martinez C, Dorronsoro M, Gonzalez CA, Key TJ, Trichopoulou A, Naska A, Vineis P, Tumino R, Krogh V, Bueno-De-Mesquita HB, Peeters PH, Berglund G, Hallmans G, LundE, Skeie G, Kaaks R, Riboli E (2003) Dietary fibre in food and protection against colorectal cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC): an observational study. Lancet 361:1496–1501

Birrenbach T, Bocker U (2004) Inflammatory bowel disease and smoking: a review of epidemiology, pathophysiology, and therapeutic implications. Inflamm Bowel Dis 10:848–859

Borgaonkar MR, Irvine EJ (2000) Quality of life measurement in gastrointestinal and liver disorders. Gut 47:444–454

Bosetti C, Levi F, Ferlay J, Garavello W, Lucchini F, Bertuccio P, Negri E, La Vecchia C (2008) Trends in oesophageal cancer incidence and mortality in Europe. Int J Cancer 122:1118–1129

Brok J, Gluud LL, Gluud C (2010) Ribavirin plus interferon versus interferon for chronic hepatitis C. Cochrane Database Syst Rev 1:CD005445

Browning JD, Szczepaniak LS, Dobbins R, Nuremberg P, Horton JD, Cohen JC, Grundy SM, Hobbs HH (2004) Prevalence of hepatic steatosis in an urban population in the United States: impact of ethnicity. Hepatology 40:1387–1395

Calkins BM (1989) A meta-analysis of the role of smoking in inflammatory bowel disease. Dig Dis Sci 34:1841–1854

Cho JH (2008) Inflammatory bowel disease: genetic and epidemiologic considerations. World J Gastroenterol 14:338–347

Chowdhury P, Gupta P (2006) Pathophysiology of alcoholic pancreatitis: an overview. World J Gastroenterol 12:7421–7427

Coons SJ, Rao S, Keininger DL, Hays RD (2000) A comparative review of generic quality-of-life instruments. Pharmacoeconomics 17:13–35

Corrao G, Tragnone A, Caprilli R, Trallori G, Papi C, Andreoli A, Di PaoloM, Riegler G, Rigo GP, Ferrau O, Mansi C, Ingrosso M, Valpiani D (1998) Risk of inflammatory bowel disease attributable to smoking, oral contraception and breastfeeding in Italy: a nationwide case-control study. Cooperative investigators of the Italian Group for the Study of the Colon and the Rectum (GISC). Int J Epidemiol 27:397–404

Cremonini F, Talley NJ (2004) Review article: the overlap between functional dyspepsia and irritable bowel syndrome – a tale of one or two disorders? Aliment Pharmacol Ther 20(Suppl 7):40–49

Danaei G, Vander HS, Lopez AD, Murray CJ, Ezzati M (2005) Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. Lancet 366:1784–1793

D'Haens G, Sandborn WJ, Feagan BG, Geboes K, Hanauer SB, Irvine EJ, Lemann M, Marteau P, Rutgeerts P, Scholmerich J, Sutherland LR (2007) A review of activity indices and efficacy end points for clinical trials of medical therapy in adults with ulcerative colitis. Gastroenterology 132:763–786

Davey-Smith G, Ebrahim S (2001) Epidemiology – is it time to call it a day? Int J Epidemiol 30:1–11

Davis GL, Alter MJ, El Serag H, Poynard T, Jennings LW (2010) Aging of hepatitis C virus (HCV)-infected persons in the United States: a multiple cohort model of HCV prevalence and disease progression. Gastroenterology 138:513–521

Drossman DA (2006) The functional gastrointestinal disorders and the Rome III process. Gastroenterology 130:1377–1390

Dufour MC (2007) Alcoholic liver disease. In: Talley NJ, Locke GR III, Saito YA (eds) GI epidemiology. Blackwell-Wiley, Oxford, pp 231–247

Eaden JA, Abrams KR, Mayberry JF (2001) The risk of colorectal cancer in ulcerative colitis: a meta-analysis. Gut 48:526–535

Editorial (no author names given) (2008) A dozen good ideas to battle hepatitis. Lancet 371:1637

Everhart JE, Ruhl CE (2009a) Burden of digestive diseases in the United States part I: overall and upper gastrointestinal diseases. Gastroenterology 136:376–386

Everhart JE, Ruhl CE (2009b) Burden of digestive diseases in the United States part II: lower gastrointestinal diseases. Gastroenterology 136:741–754

Everhart JE, Ruhl CE (2009c) Burden of digestive diseases in the United States part III: liver, biliary tract, and pancreas. Gastroenterology 136:1134–1144

Frank DN, Pace NR (2008) Gastrointestinal microbiology enters the metagenomics era. Curr Opin Gastroenterol 24:4–10

Friedenreich C, Norat T, Steindorf K, Boutron-Ruault MC, Pischon T, Mazuir M, Clavel-Chapelon F, Linseisen J, Boeing H, Bergman M, Johnsen NF, Tjonneland A, Overvad K, Mendez M, Quiros JR, Martinez C, Dorronsoro M, Navarro C, Gurrea AB, Bingham S, Khaw KT, Allen N, Key T, Trichopoulou A, Trichopoulos D, Orfanou N, Krogh V, Palli D, Tumino R, Panico S, Vineis P, Bueno-De-Mesquita HB, Peeters PH, Monninkhof E, Berglund G, Manjer J, Ferrari P, Slimani N, Kaaks R, Riboli E (2006) Physical activity and risk of colon and rectal cancers: the European prospective investigation into cancer and nutrition. Cancer Epidemiol Biomarkers Prev 15:2398–2407

Frisch M, Pedersen BV, Andersson RE (2009) Appendicitis, mesenteric lymphadenitis, and subsequent risk of ulcerative colitis: cohort studies in Sweden and Denmark. BMJ 338:b716

Gaya DR, Russell RK, Nimmo ER, Satsangi J (2006) New genes in inflammatory bowel disease: lessons for complex diseases? Lancet 367:1271–1284

Glise H, Wiklund I (2002) Health-related quality of life and gastrointestinal disease. J Gastroenterol Hepatol 17(Suppl 1):S72-S84

Goodwin CS, Mendall MM, Northfield TC (1997) Helicobacter pylori infection. Lancet 349: 265–269

Gregorio GV, Gonzales ML, Dans LF, Martinez EG (2009) Polymer-based oral rehydration solution for treating acute watery diarrhoea. Cochrane Database Syst Rev 2:CD006519

Guyatt G, Mitchell A, Irvine EJ, Singer J, Williams N, Goodacre R, Tompkins C (1989) A new measure of health status for clinical trials in inflammatory bowel disease. Gastroenterology 96:804–810

Haber P, Wilson J, Apte M, Korsten M, Pirola R (1995) Individual susceptibility to alcoholic pancreatitis: still an enigma. J Lab Clin Med 125:305–312

Hainaut P, Boyle P (2008) Curbing the liver cancer epidemic in Africa. Lancet 371:367–368

Harris HE, Ramsay ME, Andrews N, Eldridge KP (2002) Clinical course of hepatitis C virus during the first decade of infection: cohort study. BMJ 324:450–453

Hart AR, Luben R, Olsen A, Tjonneland A, Linseisen J, Nagel G, Berglund G, Lindgren S, Grip O, Key T, Appleby P, Bergmann MM, Boeing H, Hallmans G, Danielsson A, Palmqvist R, Sjodin H, Hagglund G, Overvad K, Palli D, Masala G, Riboli E, Kennedy H, Welch A, Khaw KT, Day N, Bingham S (2008) Diet in the aetiology of ulcerative colitis: a European prospective cohort study. Digestion 77:57–64

Hartling L, Bellemare S, Wiebe N, Russell K, Klassen TP, Craig W (2006) Oral versus intravenous rehydration for treating dehydration due to gastroenteritis in children. Cochrane Database Syst Rev 3:CD004390

Hawkey CJ (2009) NSAIDs and aspirin: notorious or FAMOUS? Lancet 374:93–94

Heaton KW (1987) Aetiology of acute appendicitis. Br Med J (Clin Res Ed) 294:1632–1633

Higgins PD, Johanson JF (2004) Epidemiology of constipation in North America: a systematic review. Am J Gastroenterol 99:750–759

Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411:599–603

Keus F, de Jong JA, Gooszen HG, van Laarhoven CJ (2006) Laparoscopic versus open cholecystectomy for patients with symptomatic cholecystolithiasis. Cochrane Database Syst Rev 4:CD006231

König HH, Ulshofer A, Gregor M, von Tirpitz C, Reinshagen M, Adler G, Leidl R (2002) Validation of the EuroQol questionnaire in patients with inflammatory bowel disease. Eur J Gastroenterol Hepatol 14:1205–1215

Koloski NA, Talley NJ, Boyce PM (2002) Epidemiology and health care seeking in the functional GI disorders: a population-based study. Am J Gastroenterol 97:2290–2299

Kratzer W, Mason RA, Kachele V (1999) Prevalence of gallstones in sonographic surveys worldwide. J Clin Ultrasound 27:1–7

Lang L (2005) Barry Marshall 2005 Nobel laureate in medicine and physiology. Gastroenterology 129:1813–1814

Levy RL, Whitehead WE, Von Korff MR, Feld AD (2000) Intergenerational transmission of gastrointestinal illness behavior. Am J Gastroenterol 95:451–456

Liangpunsakul S, Chalasani N (2005) Unexplained elevations in alanine aminotransferase in individuals with the metabolic syndrome: results from the third National Health and Nutrition Survey (NHANES III). Am J Med Sci 329:111–116

Loftus EV Jr (2004) Clinical epidemiology of inflammatory bowel disease: incidence, prevalence, and environmental influences. Gastroenterology 126:1504–1517

Mandayam S, Jamal MM, Morgan TR (2004) Epidemiology of alcoholic liver disease. Semin Liver Dis 24:217–232

Mangia A, Andriulli A (2010) Tailoring the length of antiviral treatment for hepatitis C. Gut 59:1–5

Marshall BJ, Warren JR (1984) Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. Lancet 1:1311–1315

Moayyedi P, Talley NJ, Fennerty MB, Vakil N (2006) Can the clinical history distinguish between organic and functional dyspepsia? JAMA 295:1566–1576

Navarro VJ, Senior JR (2006) Drug-related hepatotoxicity. N Engl J Med 354:731–739

Norat T, Bingham S, Ferrari P, Slimani N, Jenab M, Mazuir M, Overvad K, Olsen A, Tjonneland A, Clavel F, Boutron-Ruault MC, Kesse E, Boeing H, Bergmann MM, Nieters A, Linseisen J, Trichopoulou A, Trichopoulos D, TountasY, Berrino F, Palli D, Panico S, Tumino R, Vineis P, Bueno-De-Mesquita HB, Peeters PH, Engeset D, Lund E, Skeie G, Ardanaz E, Gonzalez C, Navarro C, Quiros JR, Sanchez MJ, Berglund G, Mattisson I, Hallmans G, Palmqvist R, Day NE, Khaw KT, Key TJ, San Joaquin M, Hemon B, Saracci R, Kaaks R, Riboli E (2005) Meat, fish, and colorectal cancer risk: the European prospective investigation into cancer and nutrition. J Natl Cancer Inst 97:906–916

O'Brian SJ, Halder SL (2007) Infection epidemiology and acute gastrointestinal infections. In: Talley NJ, Locke GR III, Saito YA (eds) GI epidemiology. Blackwell-Wiley, Oxford, pp 92–96

Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karaliuskas R, Duerr RH, Achkar JP, Brant SR, Bayless TM, Kirschner BS, Hanauer SB, Nunez G, Cho JH (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 411:603–606

Parashar UD, Burton A, Lanata C, Boschi-Pinto C, Shibuya K, Steele D, Birmingham M, Glass RI (2009) Global mortality associated with rota virus disease among children in 2004. J Infect Dis 200(Suppl 1):S9–S15

Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, Baker CC, Di FV, Howcroft TK, Karp RW, Lunsford RD, Wellington CR, Belachew T, Wright M, Giblin C, David H, Mills M, Salomon R, Mullins C, Akolkar B, Begg L, Davis C, Grandison L, Humble M, Khalsa J, Little AR, Peavy H, Pontzer C, Portnoy M, Sayre MH, Starke-Reed P, Zakhari S, Read J, Watson B, Guyer M (2009) The NIH human microbiome project. Genome Res 19:2317–2323

Pokorski RJ (2001) Long-term morbidity and mortality risk in Japanese insurance applicants with chronic hepatitis C virus infection. J Insur Med 33:12–36

Poynard T, Bedossa P, Opolon P (1997) Natural history of liver fibrosis progression in patients with chronic hepatitis C. The OBSVIRC, METAVIR, CLINIVIR, and DOSVIRC groups. Lancet 349:825–832

Rose J, Hawthorn RL, Watts B, Singer ME (2009) Public health impact and cost effectiveness of mass vaccination with live attenuated human rota virus vaccine (RIX4414) in India: model based analysis. BMJ 339:b3653

Rothenbacher D, Bode G, Berg G, Knayer U, Gonser T, Adler G, Brenner H (1999) Helicobacter pylori among preschool children and their parents: evidence of parent-child transmission. J Infect Dis 179:398–402

Rutgeerts P, Vermeire S, Van Assche G (2007) Mucosal healing in inflammatory bowel disease: impossible ideal or therapeutic target? Gut 56:453–455

Sandborn WJ, Feagan BG, Hanauer SB, Lochs H, Lofberg R, Modigliani R, Present DH, Rutgeerts P, Schölmerich J, Stange EF, Sutherland LR (2002) A review of activity indices and efficacy endpoints for clinical trials of medical therapy in adults with Crohn's disease. Gastroenterology 122:512–530

Santosham M, Chandran A, Fitzwater S, Fischer-Walker C, HBaqui A, Black R (2010) Progress and barriers for the control of diarrhoeal disease. Lancet 376:63–67

Sauerland S, Lefering R, Neugebauer EA (2004) Laparoscopic versus open surgery for suspected appendicitis. Cochrane Database Syst Rev 4:CD001546

Schneeweiss S, Solomon DH, Wang PS, Rassen J, Brookhart MA (2006) Simultaneous assessment of short-term gastrointestinal benefits and cardiovascular risks of selective cyclooxygenase 2 inhibitors and nonselective nonsteroidal antiinflammatory drugs: an instrumental variable analysis. Arthritis Rheum 54:3390–3398

Sonnenberg A (1995) Temporal trends and geographical variations of pepticulcer disease. Aliment Pharmacol Ther 9(Suppl 2):3–12

Sonnenberg A (2007) Time trends of ulcer mortality in Europe. Gastroenterology 132:2320–2327

Spiegel B, Harris L, Lucak S, Mayer E, Naliboff B, Bolus R, Esrailian E, Chey WD, Lembo A, Karsan H, Tillisch K, Dulai G, Talley J, Chang L (2009) Developing valid and reliable health utilities in irritable bowel syndrome: results from the IBS PROOF cohort. Am J Gastroenterol 104:1984–1991

Sprangers MA, Cull A, Bjordal K, Groenvold M, Aaronson NK (1993) The European Organization for Research and Treatment of Cancer. Approach to quality of life assessment: guidelines for developing questionnaire modules. EORTC Study Group on Quality of Life. Quol Life Res 2:287–295

Stark RG, Reitmeir P, Leidl R, König HH (2010) Validity, reliability, and responsiveness of the EQ-5D in inflammatory bowel disease in Germany. Inflamm Bowel Dis 16:42–51

Steevens J, Schouten LJ, Goldbohm RA, van den Brandt PA (2010) Alcohol consumption, cigarette smoking and risk of subtypes of oesophageal and gastric cancer: a prospective cohort study. Gut 59:39–48

Steinfeld S, Bjorke PA (2002) Results from a patient survey to assess gastrointestinal burden of non-steroidal anti-inflammatory drug therapy contrasted with a review of data from EVA to determine satisfaction with rofecoxib. Rheumatology (Oxford) 41(Supp 1):23–27

Susser M (1982) Period effects, generation effects and age effects in pepticulcer mortality. J Chronic Dis 35:29–40

Talley NJ, Locke GR III, Saito YA (eds) (2007) GI epidemiology. Blackwell-Wiley, Oxford

Thein HH, Yi Q, Heathcote EJ, Krahn MD (2009) Prognosis of hepatitis C virus-infected Canadian post-transfusion compensation claimant cohort. J Viral Hepat 16:802–813

Timmer A (2003) Environmental influences on inflammatory bowel disease manifestations. Lessons from epidemiology. Dig Dis 21:91–104

Vakil NB, Talley NJ (2007) Dyspepsia. In: Talley NJ, Locke GR III, Saito YA (eds) GI epidemiology. Blackwell-Wiley, Oxford, pp 143–148

van Duijnhoven FJ, Bueno-De-Mesquita HB, Ferrari P, Jenab M, Boshuizen HC, Ros MM, Casagrande C, Tjonneland A, Olsen A, Overvad K, Thorlacius-Ussing O, Clavel-Chapelon F, Boutron-Ruault MC, Morois S, Kaaks R, Linseisen J, Boeing H, Nothlings U, Trichopoulou A, Trichopoulos D, Misirli G, Palli D, Sieri S, Panico S, Tumino R, Vineis P, Peeters PH, van Gils CH, Ocke MC, Lund E, Engeset D, Skeie G, Suarez LR, Gonzalez CA, Sanchez MJ, Dorronsoro M, Navarro C, Barricarte A, Berglund G, Manjer J, Hallmans G, Palmqvist R, Bingham SA, Khaw KT, Key TJ, Allen NE, Boffetta P, Slimani N, Rinaldi S, Gallo V, Norat T, Riboli E (2009) Fruit, vegetables, and colorectal cancer risk: the European Prospective Investigation into Cancer and Nutrition. Am J Clin Nutr 89:1441–1452

Vandvik PO, Aabakken L, Farup PG (2004) Diagnosing irritable bowel syndrome: poor agreement between general practitioners and the Rome II criteria. Scand J Gastroenterol 39: 448–453

Völzke H, Baumeister SE, Alte D, Hoffmann W, Schwahn C, Simon P, John U, Lerch MM (2005) Independent risk factors for gallstone formation in a region with high cholelithiasis prevalence. Digestion 71:97–105

Walcher T, Haenle MM, Kron M, Hay B, Mason RA, von Schmiesing AF, Imhof A, Koenig W, Kern P, Boehm BO, Kratzer W (2005) Pregnancy is not a risk factor for gallstone disease: results of a randomly selected population sample. World J Gastroenterol 11:6800–6806

Wardlaw T, Salama P, Brocklehurst C, Chopra M, Mason E (2010) Diarrhoea: why children are still dying and what can be done. Lancet 375:870–872

Ware JE Jr, Sherbourne CD (1992) The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. Med Care 30:473–483

Ware J Jr, Kosinski M, Keller SD (1996) A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. Med Care 34:220–233

Waterland RA (2006) Epigenetic mechanisms and gastrointestinal development. J Pediatr 149:S137–S142

Wiklund I (2007) Patient-reported outcomes. In: Talley NJ, Locke GR III, Saito YA (eds) GI epidemiology. Blackwell-Wiley, Oxford, pp 24–29

Williams JG, Roberts SE, Ali MF, Cheung WY, Cohen DR, Demery G, Edwards A, Greer M, Hellier MD, Hutchings HA, Ip B, Longo MF, Russell IT, Snooks HA, Williams JC (2007) Gastroenterology services in the UK. The burden of disease, and the organisation and delivery of services for gastrointestinal and liver disorders: a review of the evidence. Gut 56(Suppl 1): 1–113

World Health Organization (2007) Global health risks: mortality and burden of disease attributable to selected major risks. WHO, Geneva

Yadav D, Whitcomb DC (2010) The role of alcohol and smoking in pancreatitis. Nat Rev Gastroenterol Hepatol 7:131–145

Web References

CDC Statistics on Viral Hepatitis (2010) (Figure) http://www.cdc.gov/hepatitis/Statistics/index.htm. Accessed 12 Mar 2011

Cochrane Colorectal Cancer Group (2009) http://cccg.cochrane.org/. Accessed 13 Mar 2011

Cochrane Hepatobiliary Diseases Group (2011) http://ctu.rh.dk/chbg. Accessed 13 Mar 2011

Cochrane IBD/FD Group (2011) http://www.cochrane.uottawa.ca/ibd/. Accessed 13 Mar 2011

Cochrane Upper Gastrointestinal and Pancreatic Diseases Group (2011) http://fhs.mcmaster.ca/ugpd/index.html. Accessed 13 Mar 2011

EORTC Website (2011) http://groups.eortc.be/qol/index.htm. Accessed 13 Mar 2011

Patient-Reported Outcome and Quality of Life Instruments Database (ProQolid)(2011) http://www.proqolid.org/. Accessed 13 Mar 2011

Rome III Diagnostic Criteria and Questionnaire (2010) http://www.romecriteria.org/. Accessed 30 May 2011

WHO Comparative Risk Assessment Project (2011) http://www.who.int/whr/2002/en/. Accessed 13 Mar 2011

WHO ICD Codes (2011) http://www.who.int/topics/classification/en/. Accessed 13 Mar 2011

Epidemiology of Psychiatric Disorders 59

Stephen L. Buka, Emma W. Viscidi, and Ezra S. Susser

Contents

S.L. Buka (✉) • E.W. Viscidi
Department of Epidemiology, Brown University, Providence, RI, USA

E.S. Susser
Mailman School of Public Health, Columbia University and New York State Psychiatric Institute, New York, NY, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_64,

59.1 Introduction

59.1.1 History of Psychiatric Epidemiology

The history of epidemiological approaches to the study of mental illness spans over 150 years. A classic paper by Dohrenwend and Dohrenwend (1982) proposed the division of psychiatric epidemiological studies into three time periods: first generation, second generation, and third generation. First-generation studies were those conducted from the end of the nineteenth century through the first half of the twentieth century. These studies relied mainly on official records and key informants to identify persons with mental disorders and, accordingly, were limited by a lack of clear operational criteria for the classification and assessment of psychiatric conditions. Although the Dohrenwends were mainly discussing prevalence studies, the historical division into these three generations is used here to describe a broader range of studies.

The first study that attempted to quantify mental illness for public health purposes was conducted in Massachusetts in 1855 by Edward Jarvis (1971). Jarvis located cases with mental disorders through the use of key informants, for instance, general practitioners and hospital records. He identified 2,632 "lunatics" and 1,087 "idiots" and analyzed the data according to demographic variables such as sex, place of birth, and economic status. A more comprehensive study was done by Arthur Mitchell in Scotland later in the nineteenth century (Susser et al. 2010).

Later in this "first generation," a landmark series of studies were led by Goldberger and Sydenstricker in the 1920s. These studies demonstrated the association between nutritional deficiency and the presence of psychosis secondary to pellagra and, moreover, went on to describe how the ecology and social inequalities of mill towns in Southern states fostered this type of nutritional deficiency (Terris 1964). In a subsequent major study of this time period, Faris and Dunham examined the geographical distribution of psychiatric patient hospitalizations in Chicago from 1922 to 1934. They found that admissions for schizophrenia were related to the ecology of urban and suburban areas, typically with higher rates reported in urban areas (Faris 1939).

During this initial period, there was also growing interest in the role of social factors in the etiology of psychiatric disorders. This idea was first introduced in the foundational work of the French scientist Émile Durkheim. In the classic text *"Le Suicide,"* published in 1897, Durkheim sought to explore the social correlates of suicide (Durkheim 1951). He explored the roles of social isolation and control among different groups and concluded that the higher rates of suicide observed among Protestants compared to Catholics and Jews and among single persons as compared to married, resulted from social isolation and deficits in social control.

In regard to prevalence, in the Dohrenwends' review of first-generation studies, they found a median prevalence rate for any psychiatric disorder of around 3.6%. These low rates, as compared to future studies, are likely due to major changes in nomenclature following World War II that broadened the definition of mental disorders, resulting in higher prevalence rates (Dohrenwend and Dohrenwend 1982).

The second generation of psychiatric epidemiological studies occurred after World War II. The community became the focus of renewed attention for both the field of psychiatry and policy makers. Several factors occurred before, during and after the war that played a role in the development of a new complement of epidemiological studies. In World War II, the diagnosis of a psychiatric disorder was the leading cause of men being rejected from military service, and after the war, more than half of all hospital beds were occupied by patients with mental conditions (Tohen et al. 2000). The National Mental Health Act of 1946 initiated a cascade of funding for psychiatric education and research that resulted in the establishment of the National Institute of Mental Health (NIMH) in 1949 and financial support for epidemiological research in service of community psychiatry.

Major improvements occurred in the second-generation psychiatric epidemiological studies: mainly (1) that study participants were interviewed directly and (2) that well-defined, large populations were studied. However, these studies continued to be limited by some of the challenges of the first-generation studies, specifically, the lack of structured diagnostic instruments and an often unclear definition of diagnostic criteria (Tohen et al. 2000). Several well-known population-based studies were conducted during this period – most notably the Midtown Manhattan (Srole et al. 1962) and Stirling County (Leighton et al. 1963) studies. Both of these used highly trained non-professional interviewers to conduct standardized interviews with study participants. This second generation of studies also began to utilize more objective measures of psychopathology as opposed to clinical judgments. The measure most often used was a 22-item questionnaire called the Langner Index developed in connection with the Midtown Manhattan Study (Langner 1962). It is widely regarded as an index of psychophysiological symptomatology and general malaise (Seiler 1973). For purposes of epidemiological studies, the questionnaire is a non-specific measure of psychological distress (Dohrenwend and Dohrenwend 1982) and did not generate psychiatric diagnoses. The Midtown Manhattan Study, initiated in 1954, had three primary objectives: to canvass the community for variations in mental health, to examine sociocultural determinants of mental health, and to establish the need for psychiatric services in the community. The Midtown Manhattan Study examined mental health in 1,660 adults culled from 1,911 Midtown dwellings, selected as a probability sample. Investigators, comprising psychiatrists and social scientists, devised a composite classification of mental health, which they called the Global Judgment of Mental Health. By and large, the same approach and set of methods employed by the Midtown Manhattan Study were used in the non-urban context of Stirling County, Nova Scotia, conducted in 1952.

These two studies reflected the recognition that the local context influenced the occurrence, expression and distribution of psychopathology – a hallmark of early epidemiological studies informing community psychiatry. The results of the Midtown Manhattan Study estimated that over 80% of those surveyed had some form of current psychopathology. However, only a quarter (23.4%) were classified as impaired, signifying the presence of marked, severe, or incapacitating symptoms – mostly anxiety. Moreover, about three quarters of those who were impaired had never sought help for their symptoms (Srole et al. 1962). In Stirling County, lifetime prevalence of any DSM-I (first edition of Diagnostic and Statistical Manual: Mental Disorders) mental disorder was about 20% (Leighton et al. 1963).

Clinical studies were also conducted during this second generation of epidemiological research. Unlike population-based studies, which surveyed community participants, these studies investigated treated populations. For example, one classic clinical study (Hollingshead and Redlich 1958) investigated the association between social class and mental illness among patients in New Haven, Connecticut, and reported higher rates of psychopathology among patients of lower social class and differences in the type of treatment received by social class. A landmark clinical study by Tsuang and Dempsey (1979) used a historical cohort design to compare patients with a diagnosed mental disorder to a surgical control group on various outcomes including patients, marital, residential, occupational, and psychiatric status. They found that, after a 30- to 40-year follow-up, patients with psychiatric disorders had worse outcomes, specifically those with schizophrenia.

The first- and second-generation studies had an important impact on the field of psychiatric epidemiology. They showed that mental illness was a substantial public health problem, that treatment service availability was lacking, and that the prevalence of psychiatric disorders varied consistently by gender, socioeconomic status, and locale (urban vs. rural) (Dohrenwend 1997). However, these studies suffered from design limitations, most notably the lack of reliable and valid assessment tools to identify cases, which impeded the quest for data on the distribution of mental disorders and potential causal factors (Tohen et al. 2000).

The third generation of epidemiological studies was marked by the development of standardized measures of mental disorders. In the late 1980s the NIMH Diagnostic Interview Schedule (DIS), a lay-administered, standardized diagnostic interview, was developed for use in the Epidemiological Catchment Area (ECA) Study (discussed in more detail below). Subsequently, a similar instrument, the Composite International Diagnostic Interview (CIDI), was used in the National Comorbidity Study (NCS) (discussed in more detail below). The creation of structured diagnostic interview instruments allowed for the reliable estimation of the prevalence of mental disorders (Tohen et al. 2000). This was the single most important contribution of the third-generation studies and resulted in the two largest population-based studies on psychiatric disorders conducted in the USA, the ECA and NCS. (Similar developments outside of the United States are discussed below.) In addition to estimates of the current prevalence of mental disorders, these instruments also collected data on lifetime prevalence and risk factors. Short-term follow-up interviews of samples from the ECA have subsequently been conducted to provide estimates of incidence (Dohrenwend 1997). These studies also examined utilization of health services, confirming prior reports that the majority of persons with mental disorders do not receive adequate treatment. The development of more accurate diagnostic instruments and symptom scales in the third generation has also resulted in better research into the natural history of psychiatric disorders (Tohen et al. 2000). For instance, the Suffolk County Study (Bromet et al. 1992) provided new insights on the time of onset of schizophrenia. And the NIMH Collaborative Depression Study focused on recovery and relapse in major depressive disorder and bipolar disorder (Keller et al. 1992, 1993).

While the above focuses on epidemiological studies in North America, similar studies have been conducted in other regions of the world. Examples include those conducted in Taiwan (Hwu et al. 1989), Brazil (Almeida-Filho et al. 1997), and India (Mehta et al. 1985). In the 1960s, the International Pilot Study of Schizophrenia (IPSS), sponsored by the World Health Organization (WHO), set out to determine if it was feasible to engage in large-scale epidemiological research of psychiatric disorders, specifically schizophrenia, with comparable methods across different contexts, and to determine what, if any, differences existed in the incidence of schizophrenia across contexts (Sartorius et al. 1974). Indeed, the IPSS, which included 1,202 participants in its initial assessment, showed that this type of study was feasible. The IPSS also demonstrated that when schizophrenia was defined as a construct with extensively tested instruments, rates were fairly consistent across contexts and nations (Sartorius et al. 1974).

Later, the WHO sponsored the Determinants of Severe Mental Disorders (DOSMeD), also known as the Ten-Country Study, which built on the work of the IPSS to study the variation in incidence and course of severe mental illness, particularly schizophrenia, across countries and cultures. The DOSMeD included sites in Denmark, India, Colombia, Ireland, the USA, Nigeria, the USSR, Japan, the United Kingdom, and the Czech Republic. The study showed that the prevalence of schizophrenia was fairly constant across countries and that schizophrenia had a better course in developing versus developed countries over a period of 2 years (Jablensky et al. 1992), which gave rise to a long debate over potential factual and artifactual explanations for these findings (Hopper and Wanderling 2000).

With recent advances in neuroscience, brain imaging, genetics, and other related fields, it is fair to say that the field of psychiatric epidemiology has, in the past decade or so, entered a new fourth generation of discovery. This includes new study designs (including family, sibling, high-risk designs, and others), new statistical methods, new diagnostic assessment techniques, and an array of new approaches to assess exposures that may reflect potential causes and mechanisms resulting in mental disorder.

59.1.2 Overview of the Chapter

In this chapter, we summarize in considerable detail what we view as both some of the major achievements of the generation of epidemiological research recently completed (e.g., the third generation of community prevalence studies) and highlight some of the exciting new approaches currently under investigation with the present generation of research. We start with a general description of the major classes of psychiatric disorders, the major classification systems used to diagnose these conditions, and summary statements regarding the global burden of mental illness. We then summarize major work in the USA and abroad on the prevalence and distribution of major psychiatric disorders and highlight two particular conditions (autism and schizophrenia) to demonstrate some of the approaches and discoveries from the current era of psychiatric epidemiology.

59.2 Overview of Major Mental Disorders

While there has been considerable change and refinement in the taxonomy and nosology of mental disorders, there is general consensus on the major categories of conditions studied in psychiatric epidemiology.

Psychotic Disorders are among the most debilitating and persistent diseases in humans. These disorders include schizophrenia, schizoaffective disorder, psychotic depression, bipolar disorder with psychotic features, and delusional disorder. Schizophrenia is the most common of the psychotic disorders and is characterized by a breakdown of thought processes and by poor emotional responsiveness. It commonly manifests as auditory hallucinations, paranoid or bizarre delusions, disorganized speech and thinking, and is accompanied by significant social or occupational dysfunction.

Mood Disorders are characterized by a disturbance in the regulation of mood, behavior, and affect. Major mood disorders include depressive disorders and bipolar disorder. One of the more common mood disorders is major depression, defined as a persistent period of dysphoric mood or loss of interest or pleasure, along with other symptoms. Bipolar disorder (also known as manic depressive illness) is a cyclical mood disorder in which episodes of major depression are interspersed with episodes of mania or hypomania.

Anxiety Disorders are disorders characterized by abnormal and pathological fear and anxiety. Anxiety disorders include generalized anxiety disorder, phobic disorders, and panic disorder. Generalized anxiety disorder is characterized by unrealistic or excessive anxiety and worry. A phobia is defined as an irrational fear that produces a conscious avoidance of the feared subject, activity, or situation. And panic disorder is a type of anxiety disorder characterized by recurring, unexpected panic attacks.

Personality Disorders are defined as inflexible and maladaptive personality traits that cause either substantial functional impairment or subjective distress. They are conceptualized as long-term characteristics of individuals that are likely to be evident by adolescence and continue throughout adulthood.

59.2.1 Refinements in the Definition and Classification of Mental Disorders: Growth of the Diagnostic and Statistical Manual of Mental Disorders (DSM) and International Classification of Diseases (ICD)

The goal of psychiatric epidemiology is to systematically study the distribution and determinants of psychiatric disorders in the population. This goal can only be obtained if there is a classification system that allows for a standardized method of collecting data on psychiatric disorders. This goal was advanced when the World Health Organization (WHO) published the sixth edition of *International Classification of Diseases* (ICD) in 1949, which included mental disorders for the first time. The American Psychiatric Association Committee on Nomenclature and Statistics developed a variant of the ICD-6 that was published in 1952, the first edition of the *Diagnostic and Statistical Manual: Mental Disorders* (DSM-I). The purpose of DSM-I was to create a common nomenclature based on a consensus of the contemporary knowledge about psychiatric disorders. These developments came in the post-World War II era (or "second generation" of psychiatric epidemiology research) to better incorporate the outpatient presentations of servicemen and veterans. DSM-I included three categories of psychopathology and was the first official manual of mental disorders to focus on clinical utility. The manual was developed to be both useful to those who diagnose and treat patients with mental illness and to provide a statistical classification to facilitate the collection of statistics on mental disorders. DSM-II was published in 1968 and had 11 major diagnostic categories and 185 diagnoses. DSM-I and DSM-II included only brief definitions of disorders which were not sufficiently operationalized to allow for reliable diagnostic judgments.

DSM-III, published in 1980, improved on previous editions by applying a descriptive approach and including diagnostic criteria. It included 265 diagnoses. The explicit diagnostic criteria in DSM-III provided the foundation for measurement of mental disorders at the population level (Eaton 2012). The emphasis on descriptive features of disorders, generally without regard for presumed etiology, allowed DSM-III to be largely embraced across disciplines. DSM-IV was published in 1994 and included 365 diagnoses. A major revision from previous versions was the inclusion of a clinical significance criterion to many of the categories, which required that symptoms cause "clinically significant distress or impairment in social, occupational, or other important areas of functioning."

A major component of psychiatric epidemiology studies is accurate psychiatric diagnosis of a clearly defined study population. Determining the prevalence of mental illness in a population requires reliable well-operationalized definitions of the disorders. There are three aspects of the DSM that have proved most valuable in the field of psychiatric epidemiology. First is the widespread acceptance of the DSM across psychiatry and related fields. This has enabled a common language for classifying mental disorders allowing for comparisons of data across studies. Second is the use of symptomatic descriptions in the DSM rather than hypothesized etiological assumptions. And third is the provision of operationalized definitions.

The introduction of DSM-III in 1980 and the subsequent development of a fully structured diagnostic interview based on it, the Diagnostic Interview Schedule (Robins et al. 1981), were critical to major population-level epidemiological studies.

One of the main questions in psychiatric epidemiology is whether the prevalence of mental disorders has been increasing over time. We have seen a clear increase in the number of persons meeting criteria for mental disorders, but this does not prove that the prevalence is increasing, as this may be due to changes in diagnostic criteria. Specifically, the DSM has expanded and the number of disorders has increased over time, which may result in an increased number of persons meeting criteria for psychiatric disorders. Researchers have questioned whether the DSM criteria are too inclusive such that we are diagnosing people with mild severity, which may not have clinical utility. This is one of the reasons that DSM-IV requires the presence of clinically significant distress or impairment to qualify for a diagnosis. As we move into the current, fourth generation of psychiatric epidemiology, there are plans for dramatic modifications to prior classification systems – both dimensional approaches based on presumed neurocognitive substrates which may cut across current diagnoses, as well as new categorical diagnoses based more on empirical data than clinical consensus.

59.2.2 Public Health Burden of Mental Disorders

Research has shown that mental disorders have a high public health burden. A global study by the World Health Organization (WHO), the so-called World Mental Health (WMH) Survey, estimated the prevalence, severity, and treatment of DSM-IV mental disorders in 14 countries (Demyttenaere et al. 2004). The findings showed a consistently high overall prevalence of mental disorders across the globe. The Global Burden of Disease (GBD) Study (Murray and Lopez 1996) found that mental disorders were responsible for 21% of the total disease burden in the world. Mental disorders ranked almost as high as cardiovascular diseases and respiratory diseases, and surpassed all different types of cancer and HIV, in Disability Adjusted Life Years (DALYs) (Ustun 1999). Specifically, depression was the fourth leading cause of disease burden, accounting for 4.4% of total DALYs (Ustun et al. 2004). Depression was reported to cause the largest amount of non-fatal burden, accounting for almost 12% of all total years lived with disability worldwide. Furthermore, WHO has projected that depression will be the second leading cause of disability by the year 2020. The DALYs are particularly high for mental disorders in part because these are non-fatal conditions of long duration, most of which have their onsets in childhood or young adulthood (Eaton 2012).

The European Study of the Epidemiology of Mental Disorders (ESEMeD) examined the impact of mental and physical disorders on work role disability and quality of life in six European countries (Alonso et al. 2004). The study showed that in each country, mental disorders resulted in loss of work and loss of quality of life.

The results suggested that mental disorders are important determinants of work role disability and quality of life, often having a greater impact than chronic physical disorders.

Studies have also shown the high cost of mental illness to individuals and society. Kessler and colleagues (2008) found that respondents with a serious mental illness in the past 12 months had earnings averaging $16,306 less than those without a mental illness, for a societal-level total of $193.2 billion. The authors used statistical modeling to show that the mean expected annual earnings of respondents with a serious mental illness in the absence of that illness were estimated at $38,851 compared to the mean observed earnings of $22,545. This difference, of approximately $16,000 (or 42%), is the estimated mean impact of serious mental illness on earnings. This study provides evidence that mental disorders are associated with substantial societal-level impairments that should be taken into consideration when making decisions about the allocation of treatment and research resources.

59.3 Summary of Findings from Major Population-Based Studies of Mental Illness

59.3.1 Major Epidemiological Studies of Psychiatric Disorder Among Adults

A number of major psychiatric epidemiological studies from the late twentieth century have helped establish methods for contemporary psychiatric epidemiology. Two classic studies in the USA include the Epidemiological Catchment Area (ECA) study and the National Comorbidity Survey (NCS). The ECA and NCS, discussed in more detail below, used reliable lay-administered structured diagnostic assessments to obtain standardized diagnostic criteria based on the DSM. The studies also compared clinical interviews with lay interviews to evaluate diagnostic validity and used advanced sampling strategies to identify nationally representative samples.

The Epidemiological Catchment Area (ECA) Study was initiated in response to the 1977 Report of the President's Commission on Mental Health, which described the state of American mental health research and services (Grob 2005). NIMH needed to provide descriptive psychiatric epidemiological data to the President's Commission, since the clinical picture of community mental health was incomplete (Regier et al 1978; Robins 1978). The ECA, therefore, sought to address the gaps identified in the President's commission report (Robins and Regier 1991).

The ECA, which commenced in 1980, was a collaborative effort between NIMH and a group of established psychiatric epidemiologists that surveyed the prevalence of mental disorders and service need and use in five US communities that had been designated Community Mental Health Center catchment areas – Baltimore, New Haven, St. Louis, Durham, and Los Angeles. Each site collected data on a common set of core questions and sample characteristics and sampled over 3,000 community residents and 500 institutionalized residents. Together, the 5-site ECA collected diagnostic service need and use data on 20,861 adults, aged 18 and over. An important innovation of the ECA was the use of a fully structured research diagnostic interview known as the Diagnostic Interview Schedule (DIS), which was shown to yield reliable and valid diagnoses of mental disorders (Helzer et al. 1985).

The ECA was the first study to document the prevalence of DSM-III disorders. While the ECA determined the epidemiological burden and service use patterns in these five communities, it also provided the indices of success of community-based treatment programs for mental illness. The ECA was critical in determining the prevalence of specific psychiatric disorders, as well as service needs and use patterns in the five communities studied. The rich information provided by the study once again supported the notion that services were inadequate relative to need; under 20% of respondents with recent mental disorders accessed services in the year prior to study participation (Robins and Regier 1991; Regier et al. 1993a). However, because the samples in the ECA were not collected to be nationally representative, there was an imperative to address epidemiological gaps regarding the prevalence and distribution of psychiatric disorders throughout the United States. Moreover, the ECA could only provide basic information regarding the comorbidity of psychiatric disorders; it was therefore necessary to determine patterns of comorbidity and the complexities of the affiliated need for services and their use.

Some of the limitations of the ECA influenced considerably the origins of the *National Comorbidity Survey (NCS)*, a truly national study of the prevalence, causes, and consequences of comorbidity between psychiatric and substance use disorders (Kessler 1994). The NCS, which began in 1990, was the first nationally representative survey of mental and substance use disorders in the United States. This study developed a new measure, the Composite International Diagnostic Interview (CIDI), to determine the prevalence and correlates of DSM-III-R (Diagnostic and Statistical Manual of Mental Disorders, third and revised edition) disorders in a nationally representative sample of 8,098 individuals, aged 18–54. The NCS was a face-to-face household survey.

The NCS found that DSM-III-R disorders are more prevalent than previously thought. The prevalence estimates of lay-administered CIDI-diagnosed psychiatric disorders were higher than those reported by the ECA, with the exception of psychotic disorders and lifetime anxiety disorders. Like the ECA, the NCS reported underuse of mental health services – about 13% of respondents accessed outpatient services in the prior 12 months. One of the most striking findings of the NCS indicated a problem with "met unneed" – that is, people with low levels of need had a higher probability of accessing treatment (Kessler et al. 1997; Mojtabai et al. 2002), with implications for policies affecting the distribution of community-based mental health services.

Similar prevalences of DSM-IV disorders and service use patterns were observed in the *NCS Replication Study (NCS-R)*, conducted about a decade later in 2001–2002 with a nationally representative sample of 9,282 respondents, aged 18–54 (Kessler et al. 2005a, b). Harkening to the results of the Midtown Manhattan Study, 26.2% of respondents reported a past-year psychiatric disorder; nearly a quarter of those with a past-year psychiatric disorder were classified as serious (Kessler et al. 2005b).

Significant unmet need for services was observed in the NCS-R; nearly 60% of those endorsing a serious past-year psychiatric disorder remained untreated (Kessler et al. 2001).

Among other large-scale epidemiological efforts, *The National Epidemiological Survey on Alcohol and Related Conditions (NESARC)* was a nationally representative United States sample of 43,093 adults aged 18 or older conducted in 2001–2002 by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) (Grant et al. 2004). This face-to-face household survey demonstrated a high degree of comorbidity between DSM-IV substance use disorders and mood and anxiety disorders.

The National Longitudinal Alcohol Epidemiological Survey (NLAES) was a nationally representative household study conducted in the United States in 1992 sponsored by NIAAA (Grant and Harford 1995). The survey consisted of face-to-face interviews with 42,862 respondents aged 18 or older. DSM-IV diagnoses were made using the Alcohol Use Disorders and Associated Disabilities Interview Schedule (AUDADIS), a fully structured psychiatric interview administered by lay interviewers. The AUDADIS was designed to determine diagnoses of alcohol abuse and dependence over the course of the previous year or over a person's lifetime, other drug use and associated disorders, and the incidence of major depression.

The Christchurch Psychiatric Epidemiology Study (Oakley-Browne et al. 1989) was conducted in 1986 as a random survey of 1,498 adults aged 18–64 years living in the Christchurch area of New Zealand. The Diagnostic Interview Schedule (DIS) was used to make DSM-III diagnoses through in-person interviews. The study used similar methods as the NIMH ECA study to allow for comparisons.

A well-known prospective, longitudinal study is the *Dunedin Multidisciplinary Health and Development Study* which has followed 1,037 babies born in Dunedin, New Zealand, between 1972 and 1973 from birth until age 38 (Silva 1990). The study has very high retention rates (96%) and has collected a wealth of information on physical and mental health across the life span, including estimates of the prevalence and incidence of psychiatric disorders from childhood through adulthood (Newman et al. 1996).

These epidemiological studies provided valuable information on the occurrence of mental disorders. All of the studies revealed that psychiatric disorders are highly prevalent in the general population, with onset often beginning in early life and with subsequent high rates of comorbidity (Insel and Fenton 2005). There is also evidence that prevalence rates may be considerably higher than reported in many prior studies. Specifically, lifetime prevalence rates have been shown to be much greater in prospective versus retrospective studies, suggesting that asking participants to recall if they have ever experienced mental disorder symptoms leads to an under-reporting (Moffitt et al. 2010). Given that most of the information about lifetime prevalence of psychiatric disorders comes from retrospective surveys, mental disorders may be far more common than previously thought (Susser and Shrout 2010).

Importantly, the high prevalence of mental disorders found in these studies led to questions about the severity of the conditions diagnosed. More recent studies have attempted to address this concern by examining the severity of diagnosed conditions.

Table 59.1 Prevalence estimates from major psychiatric epidemiology surveys in adult populations

	Study							
	ECA	NCS	NCS-R	Christchurch	NESARC	NLAES	Dunedin	World Mental Health Surveys
Any Disorder								
12 months	28.1	29.5	26.2	31.3	–	–	–	–
Lifetime	43.7	48.0	46.4	50.7	–	–	–	–
Anxiety Disorders								
12 months	11.8	17.2	18.1	9.1	11.1	–	22.8	2.4–18.2
Lifetime	19.2	24.9	28.8	10.5	–	–	49.5	4.8–31.0
Mood Disorders								
12 months	10.1	11.3	9.5	10.4	9.2	3.3	16.7	0.8–19.6
Lifetime	14.9	19.3	20.8	14.7	–	9.9	41.4	3.3–21.4
Psychotic Disorders								
12 months	1.1	0.5	–	0.2	–	–	–	–
Lifetime	1.4	0.7	–	0.4	–	–	–	–

Abbreviations: ECA, Epidemiological Catchment Area Study; NCS, National Comorbidity Survey; NCS-R, National Comorbidity Replication Survey; Christchurch, Christchurch Psychiatric Epidemiology Study; NESARC, National Epidemiological Survey on Alcohol and Related Conditions; NLAES, National Longitudinal Alcohol Epidemiological Survey; Dunedin, Dunedin Multidisciplinary Health and Development Study.

Notes: NLAES mood disorder includes major depression only; Dunedin mood disorder includes depression only. ECA and Christchurch psychotic disorders refer to schizophrenia/schizophreniform disorders only.

The values presented for the World Mental Health Surveys represent the range of prevalence estimates across the 14 countries studied.

For example, findings from the NCS-R showed that 22.3% of individuals with a disorder in the past 12 months were reported to be “serious,” 37.3% were reported to be “moderate,” and 40.4% mild (Kessler et al. 2005b). These results suggest that over half of diagnosed cases of mental illness are moderate or severe, but a substantial portion are mild.

59.3.2 Prevalence and Incidence of Major Adult Psychiatric Disorders

The major epidemiological studies discussed above have provided a wealth of information on the prevalence and incidence of psychiatric disorders in the general adult population. While recent developments in the field have moved far ahead in insights regarding risk factors, causes, and mechanisms related to the etiology of these disorders, it is valuable to summarize some of these foundational elements of the field of psychiatric epidemiology. Table 59.1 provides a summary of the prevalence estimates described below – some of the major products of the "third generation" of psychiatric epidemiology research.

Any Disorder The prevalence of any psychiatric disorder provides information on the overall burden of mental disease in the population. Prevalence estimates have been consistently high across studies. The ECA reported a 12-month prevalence of 28.1%. The NCS and NCS-R reported similar 12-month estimates at 29.5% and 26.2%, respectively. The Christchurch study reported somewhat higher rates with 31.3% of the respondents having any disorder in the past 12 months. Lifetime prevalence of any disorder is also high, with approximately half of the respondents reporting a lifetime disorder (48% in the NCS, 46.4% in the NCS-R, and 50.7% in the Christchurch study). The NCS-R also found that psychiatric disorders generally had an early onset with half of all cases reporting onset by age 14 and three quarters by age 24.

Anxiety Disorders Anxiety disorders are the most common group of disorders to occur in the general population. The ECA, NCS, and NCS-R all showed anxiety disorders to be more prevalent than mood disorders. The ECA found that 11.8% of the respondents had an anxiety disorder diagnosis in the past 12 months. Estimates in the NCS and NCS-R were similar with 12-month rates of 17.2% and 18.1%, respectively. The Christchurch study reported somewhat lower estimates of 9.1% for past 12-month diagnoses, and the NESARC found that 11.1% of the respondents had an anxiety disorder in the past 12 months. The lifetime prevalence of anxiety disorders is around one quarter. The ECA found that lifetime diagnoses of anxiety disorders occurred in 19.2% of the respondents, the NCS found a lifetime rate of 24.9%, and the NCS-R 28.8%. The Christchurch study reported a lower estimated rate of 10.5%.

Phobias have been found to be the most common anxiety disorder. The NCS-R reported that 8.7% of the respondents reported a phobia in the past 12 months with social phobia being the most common type. The NCS-R also reported on the prevalence of panic attacks, a subtype of anxiety disorder, with a lifetime prevalence of 28.3% and 12-month prevalence of 11.2% (Kessler et al. 2006). This rate is somewhat higher than that of the NCS (7.3%).

Mood Disorders Mood disorders are the second most prevalent class of disorders. The ECA found that 10.1% of the respondents had a mood disorder diagnosis in the past 12 months. Estimates in the NCS, NCS-R, and Christchurch study were comparable at 11.3%, 9.5%, and 10.4%, respectively. The NESARC found that 9.2% of the respondents had a mood disorder in the past 12 months, while the NLAES found a considerably lower prevalence of 3.3% (which is likely due to the fact that they only measured major depression).

The ECA found that lifetime diagnoses of mood disorders occurred in 14.9% of the respondents, while estimates in the NCS and NCS-R were higher at 19.3% and 20.8%, respectively. The Christchurch study found a prevalence of 14.7% for lifetime diagnosis, and the NLAES reported that 9.9% of the respondents had a lifetime diagnosis.

Psychotic Disorders Psychotic disorders are less common than mood and anxiety disorders and are less often studied in epidemiological surveys of the general population. The ECA found a lifetime prevalence of psychotic disorders of 1.4%, while the NCS reported lower estimates at 0.7% lifetime and 0.5% in the past 12 months. The Christchurch estimates were even lower at 0.2% lifetime and 0.4% past 12 months (which is likely explained by the fact that only schizophrenia/schizophreniform disorders were diagnosed).

Comorbidity of Psychiatric Disorders The ECA was among the first studies to show how common psychiatric comorbidity of psychiatric disorders is in the general population. Over 54% of the ECA respondents with a lifetime history of a disorder were found to have a second diagnosis. Similarly the NCS found that over 50% of all lifetime disorders occurred in a small proportion of the respondents with a history of three or more comorbid psychiatric disorders (Kessler et al. 1994). And the NCS-R also found high rates of psychiatric comorbidity; nearly 30% of the people surveyed had two disorders, and almost 20% had three (Kessler et al. 2005a, b).

Sociodemographic Risk Factors for Adult Psychiatric Disorders Studies have consistently shown that women have higher rates of mood and anxiety disorders, while men have higher rates of substance use disorders (Wells et al. 1989; Kessler et al. 1994). The prevalence of most disorders has been shown to decline with age and higher socioeconomic status (Regier et al. 1993b; Kessler et al. 1994). A family history of mental illness consistently elevates the risk for psychiatric disorder (Kendler et al. 1997).

Prior studies have also shown that persons who are separated or divorced are more likely to have a psychiatric disorder than those who are married (Regier et al. 1993b; Bjorkenstam et al. 2012). For instance, the National Survey of Mental Health and Wellbeing of Adults, a random sample of adults in Australia, found that married people were the least likely to suffer from a mental disorder, while divorced and separated adults were the most prone to mood and anxiety disorders and never-married adults were the most at risk for drug and alcohol disorders (Andrews et al. 1999). The nature of this association is complicated and may vary by sex and type of mental disorder (Fox 1980). It remains unclear whether the protective effect of marriage is due to the social role of being married (social causation) or characteristics of the individuals who get married (social selection). Persons with mental disorders may be less likely to marry and more likely to divorce; and separation and divorce or the death of a spouse may adversely affect a person's mental health. A study by Bjorkenstam et al. (2012) supported both the selection hypothesis, linking healthy individuals to long and stable marriages, and the social causation hypothesis, linking the stress of recent divorce to increased psychiatric disorder for both women and men. In a study of women, Afifi et al. (2006) found that never-married mothers from the NCS were similar to married mothers in lifetime prevalence rates of mental disorders. However, separated/divorced mothers had increased odds of various disorders such as depression and anxiety, as compared to married mothers.

59.3.3 Major Epidemiological Studies of Children and Adolescents

There has also been substantial epidemiological research on the occurrence of mental disorders among children and adolescents. The global burden of mental illness in young people is significant; the WHO predicts that by 2020, childhood neuropsychiatric disorders will become one of the leading causes of morbidity, mortality, and disability among children worldwide (USDHHS 1999). This age group is also important in light of the findings that many psychiatric disorders have their onset in early life (Kessler et al. 2007). Most major epidemiological studies, such as the National Comorbidity Survey Replication (NCS-R), indicate that many individuals with mental illness report onset of their disorders in adolescence or early adulthood. Mental disorders, particularly if left untreated, are likely to persist into adulthood. According to the 2001 Surgeon General's Conference on Children's Mental Health, approximately 74% of 21-year olds with mental disorders had prior mental health problems. Furthermore, researchers are interested in examining early risk and protective factors for psychiatric disorders in the hopes that early intervention can reduce the burden of disease in children, adolescents, and adults. Considerable scientific ground has been covered regarding the epidemiological study of child and adolescent psychopathology in the past 30 years. Most notably, much more is now known about the measurement, community study, prevalence, and risk factors for psychiatric disorders, particularly those of older children and adolescents. Epidemiological data suggests that roughly 20% of children aged 1–18 years are in need of mental health services (Burns et al. 1995; Shaffer et al. 1996).

A number of important epidemiological studies have been conducted in child and adolescent populations. One of the first major studies was the *Isle of Wight/Inner London Borough Study* (Rutter et al. 1975). The study reported on the prevalence of psychiatric disorders, measured by parent interviews, in a sample of 10-year-old children from inner city London ($n = 1{,}689$) and the Isle of Wight ($n = 1,279$). *The Ontario Children's Study* (Offord et al. 1987) studied the 6-month prevalence of four child psychiatric disorders (conduct disorder, hyperactivity, emotional disorder, and somatization) among approximately 2,670 children 4–16 years of age residing in Ontario, Canada.

The Great Smoky Mountains Study (GSMS) was a population-based study of 1,420 children aged 9–13 years in North Carolina that began in 1992 (Costello et al. 2003). The study was a longitudinal design in which three cohorts of children were examined at age 9, 11, and 13 years at intake. The children were assessed annually through age 16 years for DSM-IV disorders using the Child and Adolescent Psychiatric Assessment (CAPA), an in-person diagnostic interview. Both parents and children were interviewed, leading to one of the major methodological challenges in the field of child psychiatric epidemiology – namely, how to treat the large number of families in which child and parent diagnostic reports are inconsistent. *The Puerto Rico Community Study* examined DSM-IV disorder in 1,886 children aged 4–17 years using the Diagnostic Interview Schedule for Children (DISC) (Canino et al. 2004).

The National Comorbidity Survey Adolescent Supplement (NCS-A) (Merikangas et al. 2009b) was a nationally representative face-to-face survey of 10,148 adolescents aged 13–18 years in the United States conducted from 2001 to 2004. The purpose of the survey was to provide the first nationally representative estimates of the prevalence, correlates, and risk and protective factors of mental disorders in US adolescents. DSM-IV diagnoses were made based on a modified version of the Composite International Diagnostic Interview (CIDI), a fully structured research diagnostic in-person interview designed for use by trained lay interviewers. In addition to in-person interviews with adolescents, there was also a parent self-report questionnaire.

59.3.4 Prevalence and Incidence of Major Child and Adolescent Psychiatric Disorders

As discussed previously, psychiatric disorders among children and adolescents are widespread. Previous summaries of the prevalence of mental disorders in population surveys of US youth conclude that about one out of every four youth meet criteria for a DSM disorder (Costello et al. 2004). A review by Merikangas et al. (2009a) found that about one third of youth experience a mental disorder in their lifetime; and fewer than half with a current disorder receive mental health treatment (Merikangas et al. 2009a). Angold and Costello (1995) summarized findings from previous studies and found that the 6-month prevalence rate for any psychiatric disorder ranged from 17% to 27%. The most common forms of psychopathology were anxiety disorders, behavior disorders, and mood disorders, in that order.

In a review of 52 studies of psychopathology in children and adolescents, Roberts et al. (1998) found that prevalence estimates of psychopathology ranged from 1% to nearly 51%, with a mean prevalence of any mental disorder of 15.8%. In a similar review by Costello et al. (2005), the median prevalence estimate of psychiatric disorders was 12%. Importantly, the authors noted that disorders that often appear first in childhood or adolescence are among those ranked highest in the World Health Organization's estimates of the global burden of disease.

The Ontario Children's Study found that among children 4–16 years of age, the 6-month prevalence rate of any disorder was 18.1%. The Quebec Child Mental Health Survey (QCMHS) was conducted in 1992 on a representative sample of 2,400 children and adolescents aged 6–14 years in Quebec (Breton et al. 1999). The 6-month prevalence of any disorder was 19.9% according to parent report and 15.8% according child report. The rates of one or more anxiety disorders ranged from 58% to 17.5%, while rates of depressive disorders ranged from 1.1% to 3.5%.

The GSMS found a 3-month prevalence of any disorder of 13.3%; but by age 16 years, 36.7% (lifetime prevalence) of children had met DSM-IV criteria for one or more disorders. The authors concluded that while the proportion of children with a diagnosis at a given time was small, almost three times this number had one or more disorders over the entire study period suggesting that cross-sectional studies may underestimate the prevalence of psychiatric disorders among children.

Furthermore, 25.5% of children with a mental disorder diagnosis had two or more conditions. There was significant comorbidity among the behavioral disorders and between anxiety and depression. In particular, there was a strong association between oppositional defiant disorder and depression, and depression and anxiety.

The Puerto Rico Community Study found that 16.4% of children aged 4–17 years had a DSM-IV diagnosis in the past year. The most prevalent disorder was attention deficit/hyperactivity disorder (ADHD) occurring in 8% of children. They also found that 6.9% of children had an anxiety disorder and 3.4% of children had a depressive disorder.

The NCS-A showed high rates of mental disorders among adolescents; 40% of adolescents had a DSM-IV disorder in the past year (Kessler et al. 2012) and approximately one in every 4–5 youth met criteria for a mental disorder with severe impairment across their lifetime (Merikangas et al. 2010). Anxiety disorders were found to be the most prevalent mental disorder (24.9% 12-month prevalence and 31.9% lifetime prevalence), followed by behavior disorders (16.3% 12-month prevalence and 19% lifetime prevalence) and mood disorders (10% 12-month prevalence and 14.3% lifetime prevalence). Comorbidity was common with approximately 40% of those with one class of disorder also meeting criteria for another disorder. The median age of onset was earliest for anxiety disorders (6 years), followed by 11 years for behavior disorders, and 13 years for mood disorders.

Table 59.2 provides a summary of the prevalence estimates described above.

Risk Factors for Child and Adolescent Psychiatric Disorders A number of risk factors for child and adolescent psychiatric disorders have been identified. Findings from twin and adoption studies have firmly established the contribution of genetics to the occurrence of disorders in children. However, the mechanism of inheritance is complex (State and Dykens 2000) and the heritability estimate of each childhood psychiatric disorder is typically considerably less than 100%. Psychopathology in parents has also consistently been shown to be associated with mental disorders in offspring (Angold and Costello 1995).

Variation in rates of mental disorders by sex has been observed in child and adolescent populations. Most studies have found that behavior disorders are more common in boys and depressive and anxiety disorders are more common in girls (Offord et al. 1987; Costello et al. 2003; Canino et al. 2004; Merikangas et al. 2010; Kessler et al. 2012). The GSMS found that overall psychiatric disorders were more prevalent in boys aged 9–16 years than in girls as a result of the much higher rate of conduct disorder and attention deficit-hyperactivity disorder (ADHD) in boys.

Sex differences in developmental trajectories for mental disorders have also been observed. For example, the Ontario Children's Study found that rates of emotional disorders in boys and girls were similar in the 4- to 11-year-old age group at around 10%. However, by ages 12–16 the rate had dropped to 4.9% in males but increased to 13.6% in females. With regard to depression, among preadolescents, studies have shown either no sex differences in rates of depression or even higher rates in preadolescent boys (Merikangas and Avenevoli 2002). During adolescence,

Table 59.2 Prevalence estimates from major psychiatric epidemiology surveys of child and adolescent populations

	Study				
	Great Smokey Mountain Study (GSMS)	Ontario Children's Health Study	Quebec Child Mental Health Survey (QCMHS)	The Puerto Rico Community Study	NCS-A
Age of sample	9–16 years	4–16 years	6–14 years	4–17 years	13–18 years
Any disorder					
3 months	13.3	–	–	–	–
6 months	–	18.1	15.8 (child) 19.9 (parent)	–	–
12 months	–	–	–	16.4	40.3
Lifetime	36.7	51.3	–	–	–
Mood disorder					
3 months	2.2	–	–	–	–
6 months	–	–	3.4 (child) 1.7 (parent)	–	–
12 months	–	–	–	3.4	10.0
Lifetime	–	–	–	–	14.3
Anxiety disorder					
3 months	3.4	–	–	–	–
6 months	–	–	9.1 (child) 14.7 (parent)	–	–
12 months	–	–	–	6.9	24.9
Lifetime	–	–	–	–	31.9
Behavior disorder					
3 months	7.0	–	–	–	–
6 months	–	–	–	–	–
12 months	–	–	–	11.1	16.3
Lifetime	–	–	–	–	19.1

Abbreviations: NCS-A, National Comorbidity Survey Adolescent Supplement.
In the QCHMS, 'child' refers to diagnosis based on child report and 'parent' refers to diagnosis based on parent report.

however, rates of depression are higher among girls than boys (Cohen et al. 1993; Kessler and Walters 1998).

Overall psychiatric disorders have been shown to become more prevalent as children age. The NCS-A found that the lifetime prevalence was 8.4% in those 13–14 years versus 15.4% in those 17–18 years. However, there is variation in age-related rates by specific disorder. For example, prevalence rates of depressive disorders and anxiety disorders have been shown to increase with age, while prevalence rates of ADHD decrease with age (Breton et al. 1999; Canino et al. 2004).

Higher prevalence rates of child mental disorders have been found in urban versus rural areas (Offord et al. 1987; Costello et al. 1996; Canino et al. 2004). This was shown in the Isle of Wight/Inner London Borough Study in which the prevalence of any psychiatric disorder was higher in the London borough (25.4%), an urban area, than the Isle of Wight (12%), a rural area. Other risk factors for psychiatric disorders include lower socioeconomic status (SES) (Rutter et al. 1975; Costello et al. 1996; Merikangas et al. 2010), having unmarried parents (Canino et al. 2004; Merikangas et al. 2010), and a variety of exposures and complications during the prenatal period (Buka et al. 1993).

59.4 The Fourth Generation of Psychiatric Epidemiology: Examples of Specific Disorders

The material described thus far summarizes some of the major accomplishments of what we characterize as the “third generation” of psychiatric epidemiology research – largely conducted between 1980 and 2000. These large-scale community studies have solidified epidemiological knowledge of the prevalence, incidence, risks, and course of major disorders of children, youth, and adults. Since the turn of the century, there has been rapid expansion in the number and variety of epidemiological investigations pursued worldwide. With recent advances in neuroscience, brain imaging, genetics, and other related fields, the field of psychiatric epidemiology has entered a new fourth generation of research. This includes more extensive and sophisticated use of certain study designs (including family, sibling, high-risk designs, nested case-control studies, and others), new statistical methods, new diagnostic assessment techniques, and an array of new approaches to assess exposures that may reflect potential causes and mechanisms resulting in mental disorder. While space does not permit us to describe the full array of exciting work conducted over the past 10–20 years, we highlight two specific disorders to give some sense of these recent developments. We highlight two of the more severe and persistent mental disorders – autism and schizophrenia.

59.4.1 Autism Spectrum Disorders

Autism spectrum disorders (ASD) are a heterogeneous group of neurodevelopmental disorders characterized by deficits in social interaction and communication and the presence of restricted and repetitive behavior (Diagnostic and Statistical Manual of Mental Disorders, fourth edition, text revision). These core characteristics are frequently accompanied by impairments in cognition, learning, attention, and sensory processing. ASDs include three main conditions classified as pervasive developmental disorders in the DSM: autistic disorder, Asperger syndrome, and pervasive developmental disorder not otherwise specified (PDD-NOS). ASDs are typically diagnosed in early childhood and remain lifelong conditions. Symptoms are usually present before 3 years of age, but most individuals are now diagnosed between 3 and 4 years of age, with boys being diagnosed, on average, earlier than girls (Chakrabarti and Fombonne 2001; Yeargin-Allsopp et al. 2003).

Autism was first described in the US and European medical literature in the 1940s. Through the 1980s, it was believed to be rare, with an estimated prevalence of 5 per 10,000 persons. The estimated prevalence has increased dramatically over the past 30 years, and at the present ASDs are estimated to occur in 110 in 10,000 children in the United States (Kogan et al. 2009; Principal Investigators 2012). Prevalence estimates in Europe are similar to the USA (Baron-Cohen et al. 2009), while prevalence rates in other parts of the world are somewhat lower. For example, a study conducted in Japan found a prevalence of 27.2 per 10,000 individuals (Honda et al. 2005).

A steady increase in ASD prevalence has been consistently shown in the USA and globally. In the USA, the prevalence of ASD in 8-year-old children increased by 78% between 2002 and 2008, with larger increases seen in children of borderline and average or above-average intellectual functioning (Principal Investigators 2012). A study of time trends in ASD diagnoses among children aged 36 months and younger in Massachusetts found that ASD incidence increased from 56 per 10,000 among the 2001 birth cohort to 93 per 10,000 in 2005 (Manning et al. 2011).

A little less than half of the considerable increase in the prevalence of ASD is best explained by broadening diagnostic criteria and increased awareness. Significant changes occurred in the diagnostic criteria from the DSM-III to DSM-IV, during which time increases in prevalence were observed (King and Bearman 2009). Furthermore, studies have shown that diagnostic substitution, which occurs when children with another diagnosis have their diagnosis changed to ASD, also explains some of the rise in ASD prevalence. One study found that 26.4% of the increased autism caseload in California was associated with diagnostic change (specifically, children previously diagnosed with only intellectual disability) (King and Bearman 2009). A large part of the increase remains unexplained, however, by any of these factors. Therefore, we must leave open the possibility that there has actually been an increase in prevalence due to a true increase in incidence.

There has been a particularly large increase in children meeting criteria for ASD who do not have cognitive delay. The inclusion of Asperger disorder in DSM-IV is likely to have contributed to the increased prevalence of ASDs by encouraging identification of high-functioning children with social problems but no language delay (McPartland et al. 2012). Furthermore, DSM-III applied a monothetic approach (i.e., an individual must meet all diagnostic criteria), while DSM-III-R used a set of polythetic criteria (i.e., individuals must meet only a subset of a range of criteria), which also broadened the diagnostic concept. A recent study revisited a sample of children diagnosed with ASD by DSM-III-R to determine how many would meet criteria under DSM-IV (Miller et al. 2013). The results showed that 91% of participants met both criteria and a large number of participants who did not meet criteria under DSM-III (59%) did meet DSM-IV-TR criteria. Similar retrospective studies that have applied current ASD diagnostic criteria to samples diagnosed in the past show that prior studies may have underestimated prevalence

(Heussler et al. 2001; Charman et al. 2009). Finally, recommendations from the American Academy Pediatrics to screen all children for ASD, and increased public awareness, have helped to improve diagnosis and treatment and may explain increases in prevalence (Duchan and Patel 2012).

Innovative epidemiological methods have been used in the field of autism epidemiology in recent years to better understand the distribution and determinants of ASD. Researchers have used new study designs and methods to obtain more accurate prevalence estimates and to attempt to explain increases in prevalence over time. Various epidemiological studies have been conducted to examine time trends in the prevalence of ASD. To study trends, repeated surveys in defined geographical areas at different points in time can be used, provided that methods are kept relatively constant (Fombonne 2009). These studies have generally shown increases in prevalence over time (Gillberg 1984; Kawamura et al. 2008). Also, age-period-cohort analyses of time trends in California have established that the increased prevalence of ASD in that state is primarily a birth cohort effect, which limits the possible explanations to some degree (Keyes et al. 2012).

Many studies have relied on public records (medical and educational) to obtain prevalence estimates in large population-based samples (Croen et al. 2002; Principal Investigators 2012). The most prominent ongoing population-based surveillance program is the Autism and Developmental Disabilities Monitoring (ADDM) Network conducted in the United States by the Centers for Disease Control (CDC) (Principal Investigators 2012). The ADDM uses a standardized protocol across various study sites in the USA to examine medical and educational records of 8-year-old children to diagnose ASD. The ADDM studies have been valuable in providing data on the prevalence of ASDs, describing the population of children with ASD, and identifying changes in prevalence over time.

Recent studies have challenged the diagnostic criteria for ASD outlined in the DSM. For example, Lord and colleagues (2012) found that clinical distinctions among categorical diagnostic subtypes of ASD (autistic disorder, Asperger disorder, and pervasive developmental disorder not otherwise specified) were not reliable across multiple study sites and concluded that dimensional descriptions of core features, along with language and cognitive function, would be more useful. Studies like this one provide support for the diagnostic changes proposed for DSM-V, which will eliminate subgroups in favor of one overall diagnosis – autism spectrum disorder.

A consistently shown risk factor for ASD is male sex. Studies show that the odds of having ASD are approximately four times greater for boys than girls (Kogan et al. 2009). However, when females are affected, they generally have lower functioning (Rivet and Matson 2011). Comorbid conditions are common in individuals with ASD. Kogan et al. (2009) found that 87.3% of children with ASD also had attention deficit disorder or attention deficit/hyperactivity disorder (ADHD), anxiety problems, behavioral or conduct problems, depression, or developmental delay. Among these conditions, the most common was ADHD, with an estimated prevalence of 47.2 per 100 among individuals with ASD. Psychiatric disorders in family members are also more frequent in individuals with ASD (Duchan and Patel 2012). Recent studies have shown that schizophrenia coaggregates with ASD in many families (Sullivan et al. 2012).

Another common condition associated with ASD is intellectual disability (ID), defined as IQ at or below 70. Current estimates are that around 40% of children with ASDs have ID (Principal Investigators 2012). Epilepsy is also known to be comorbid with ASD and occurs in approximately 14% to 16% of individuals with ASD (Tuchman et al. 1991; Levy et al. 2010), a greater proportion of whom also have ID. Longitudinal studies that follow persons with ASD into adulthood have reported higher epilepsy rates (Danielsson et al. 2005). Genetic and medical problems are also common; one study found that 15% of individuals with pervasive developmental disorder had a known genetic disorder (Chudley et al. 1998).

Outcomes in persons with ASD vary greatly depending on the type of ASD, severity of symptoms, and presence of comorbid conditions. Individuals with ASD have higher mortality rates as compared to the general population (Gillberg et al. 2010), in particular persons with ASD and epilepsy (Pickett et al. 2011). While ASD can improve through treatment, in particular through behavioral therapies implemented in early childhood (Smith et al. 2000; Dawson et al. 2010), for most individuals the condition remains a significant lifelong impairment. Much of the epidemiological research on autism has focused on children. Very little is known about adults with ASD. However, it is critical to know more about this population because, at present, many adults with ASD were misdiagnosed or undiagnosed as children and there is a large number of children with autism who will soon become adults (Bresnahan et al. 2009). Furthermore, addressing the needs of adults with autism is a major public health concern. As the number of adults with ASD increases, there is a growing need for expanded services and an improved understanding of the types of services and treatment needed for adults. Epidemiologists must employ a life course approach to autism, which will improve our understanding of the disorder (Bresnahan et al. 2009).

The use of new study designs has moved the field of autism epidemiology forward. Infant sibling studies have played an important role in ASD research over the past decade. These studies have provided important new knowledge about the early signs of ASD, have improved our understanding of the developmental course of autism, and are useful methods to identify risk factors for autism. The use of infant sibling designs is effective because of the increased recurrence rate of autism in siblings and the opportunity to observe early behavioral markers of autism (Newschaffer et al. 2012). Furthermore, the enrollment of participants during pregnancy allows for prospective research on risk factors occurring during the perinatal period. This type of study has been labeled as an enriched-risk pregnancy cohort study and has been applied in studies such as the Early Autism Risk Longitudinal Investigation (EARLI) (Newschaffer et al. 2012). Large birth cohorts have also been employed in the research of ASD. For example, the Autism Birth Cohort (ABC) identifies cases of autism through population screening of participants from the Norwegian Mother and Child Cohort, a large, randomly selected birth cohort in Norway (Stoltenberg et al. 2010). More than 107,000 children have been screened for autism. Cases are compared to controls from the same cohort in a nested case-control design to identify genetic and environmental risk factors for autism.

There is a clear genetic risk for ASD as evidenced by twin and family studies. The relative risk of ASD is 22 times greater in individuals who have a sibling with ASD autism (Lauritsen et al. 2005), and there is a higher concordance rate in monozygotic (MZ) than dizygotic (DZ) twins (Lichtenstein et al. 2010). Based on twin studies, the heritability of ASD is estimated to be between 60% and 90%, making autism one of the most genetic of all developmental neuropsychiatric disorders (Kumar and Christian 2009). While prior studies showed significantly higher concordance rates in MZ twins (Steffenburg et al. 1989; Bailey et al. 1995), more recent studies have provided evidence of a greater role for environmental factors. Rosenberg and colleagues (2009) found a higher concordance rate in DZ twins than previously reported (31%), and Hallmayer et al. (2011) found a rate of concordance in DZ twins of 31% to 36%, concluding that environmental factors explain about 55% of the liability to autism. These studies have increased interest in research on environmental risk factors for ASD, in particular those occurring during the prenatal period. Environmental risk factors that have been identified include perinatal complications (Hultman et al. 2002; Gardener et al. 2009) and increased maternal (Manning et al. 2011) and paternal age (Durkin et al. 2008). In a meta-analysis, Gardener and colleagues (2009) reported that advanced parental age at birth, maternal prenatal medication use, bleeding, gestational diabetes, being firstborn, and having a mother born abroad were associated with increased risk of autism. Durkin et al. (2008) showed that both maternal and paternal age were independently associated with autism.

The recent interest in environmental risk factors for ASD has resulted in the identification of potential protective factors that reduce ASD risk. For example, in a Norwegian cohort of mothers and children, maternal use of folic acid supplements in early pregnancy was associated with a reduced risk of severe language delay, a condition associated with ASD, in children at age 3 years (Roth et al. 2011). And Schmidt and colleagues (2011) found that mothers of children with autism were less likely than those of typically developing children to report having taken prenatal vitamins before or in early pregnancy. The use of folic acid supplements and prenatal vitamins may reduce the risk of having a child with autism.

Exciting research investigating genetic risk factors for ASD has also been conducted in recent years. The advent of microarray technology has led to a revolution in the discovery of copy number variants (CNVs) in autism. A number of CNVs have been identified that occur more frequently in persons with ASD than controls (Morrow 2010). Curiously, many of these same CNVs are also related to other psychiatric disorders, such as schizophrenia (Sanders et al. 2012). Recently, genome-wide association studies (GWAS) of autism have been conducted that have found chromosomal deletions (Kumar et al. 2008) and copy number variants (Marshall et al. 2008) associated with autism. Currently ongoing are “next-generation” genomic studies, which use approaches such as whole genome sequencing that are far more comprehensive than GWAS approaches. One of the earliest to be published showed that paternal age is associated with increased risk of de novo SNP mutations (not just CNVs) and suggested that some of these mutations were related to autism and/or schizophrenia (Kong et al. 2012). While these studies have yielded important findings, they have also underscored the complexity of autism inheritance.

Finally, recent advances in neuroscience and brain imaging have resulted in a better understanding of the brain development of persons with ASD. For example, a longitudinal study of brain growth by Schumann and colleagues (2010) found an abnormal growth rate of multiple areas of the brain among children with autism. The widespread use of standardized diagnostic instruments in the autism epidemiology field, in particular the Autism Diagnostic Interview Revised (ADI-R) (Rutter et al. 2003) and the Autism Diagnostic Observation Schedule (ADOS) (Lord et al. 2000) has allowed for better ascertainment of cases and the ability to make comparisons across studies. As DSM criteria for ASD continue to change (Wing et al. 2011), there will remain challenges in studying the incidence, prevalence, and risks for the disorder. While several risk factors have been identified, the exact cause or causes of ASD are still unknown. Further research on risk factors for ASD is needed, in particular those that occur during the perinatal period or which may be involved in the interplay of genes and environment. There also continues to be a need for studies examining the diagnostic criteria for ASD to better refine case identification.

59.4.2 Schizophrenia

Schizophrenia is a psychotic disorder characterized by a breakdown of thought processes and poor emotional responsiveness. The essential features are the presence of specific psychotic symptoms (delusions, hallucinations, etc.), social or occupational dysfunction, and a 6-month duration of illness (Tsuang and Tohen 2002).

In a review of 188 studies, McGrath et al. (2008) found a median point prevalence of schizophrenia of 4.6 per 1,000 and lifetime prevalence of 7.2 per 1,000. The risk of developing the illness over the life course was 0.7%. The Dutch national morbidity survey reported a lifetime prevalence of 3.7 per 1,000 (van Os et al. 2001).

Several studies have estimated the incidence of schizophrenia. A systematic review of 158 incidence studies reported a median annual incidence of 15.2 per 100,000 (McGrath et al. 2008). In a global study (WHO ten-country study, Jablensky et al. 1992), the estimated annual incidence of schizophrenia ranged from 16 to 40 per 100,000 using broad diagnostic criteria and 7 to 14 per 100,000 using narrow criteria. Eaton et al. (1998) reviewed incidence studies and reported a median annual incidence rate of 0.20 per 1,000. Currently many investigators question whether narrowly diagnosed schizophrenia is clearly separable from other psychotic disorders, at least in terms of genetic and early environmental causes. The overall incidence of psychoses varies, of course, but most studies suggest that the proportion of people who develop psychoses in high-income countries is similar to the proportion who develop ASD. Studies of incidence and prevalence for autism and schizophrenia from low-and middle-income countries (LMIC) are not yet sufficient to draw general conclusions that pertain beyond high-income countries, although LMIC represent ~90% of the world's population. With the growth of research capacity in LMIC, recent pioneering studies of incidence (Menezes et al. 2007) and prevalence (Kebede et al. 2003; Kim et al. 2011; Phillips et al. 2004) are now emerging for both ASD and schizophrenia, and we should soon have enough data to paint a broader, global picture for these disorders.

Although schizophrenia has been studied over the past century, its causes remain largely unknown. Among the reasons are the inherent etiological heterogeneity of schizophrenia as a "syndrome" defined by a confluence of symptoms, and the related historical practice of assigning a different diagnosis than schizophrenia once a specific cause of the syndrome is identified (e.g., pellagra, syphilis, drug-induced psychosis). There are clearly both environmental and genetic causes, and the interplay among these causes is likely to be important.

Schizophrenia has been shown to run in families (Jones and Cannon 1998; Gottesman and Shields 1982); heritability estimates range widely with an upper bound of ~80% (Sullivan et al. 2003; Gottesman and Shields 1982). Individuals are at higher risk if a biological parent or sibling had the disease, and intriguingly, the increased risks related to parents and siblings are somewhat different despite the fact that parents and siblings share about 50% of their genes with an index individual (Heston 1966; Jones and Cannon 1998; Gottesman and Shields 1982). Twin studies show a substantially greater concordance in monozygotic as compared to dizygotic twins (Cannon et al. 1998; Sullivan et al. 2003).

In recent years, major advances have been made in identifying specific genetic mutations related to schizophrenia. About 5 years ago, we argued that individually rare mutations would turn out to be important causes (McClellan et al. 2007), and since then, a large number of studies have been published that identify rare mutations as strong causes (e.g., Gilman et al. 2012). However, it is still not known whether these individually rare mutations will be related to a very large proportion of cases. Thus, the counterargument that schizophrenia is related to multiple common genetic variants, each with a weak effect, could also be true for a proportion of cases (Owen et al. 2009). A very large number of common polymorphisms have been related to schizophrenia in GWAS studies (with weak effects as would be expected). Overall, with a few notable exceptions (e.g., 22q deletions), results across studies are not yet consistent enough to draw definitive conclusions about specific genes or gene regions, but do suggest that both rare mutations and common variants within pathways relating to neuronal (especially synaptic) development play a role. An intriguing recent paper reports that genes within these pathways are important for both ASD and schizophrenia and that there is substantial overlap for these two disorders (Xu et al. 2012). Finally, we note that genetic studies of both rare and common genetic variants are frequently subject to potential bias due to overlooking basic epidemiological precepts (e.g., Schwartz and Susser 2010), although we do not think such potential bias has altered the overall pattern of results described here.

A number of environmental risk factors for schizophrenia have also been identified. With regard to demographic factors, one of the most consistent findings is the lower socioeconomic level of persons with schizophrenia compared to unaffected individuals A large body of research has addressed various hypotheses for this association, ranging from theories of social selection/social drift in which individuals who have or are prone to schizophrenia "drift" progressively downwards in social class as a result of the disorder and/or are prevented from attaining higher social class levels, to theories of social causation involving environmental conditions occurring among persons of lower SES. Both earlier (Dohrenwend et al. 1992) and recent cohort studies tend (Corcoran et al. 2009) to support the social selection view, with only children of the lowest social classes at elevated risk to develop schizophrenia. By contrast, there has been a large recent body of work indicating that ethnic minorities migrating to developed countries are at increased risk for developing schizophrenia and psychotic symptomatology, due to presumed (and yet unexplained) mechanisms of social causation (Morgan and Fearon 2007; Veling et al. 2011). There is also evidence that schizophrenia is more common among males than females (McGrath et al. 2004). McGrath et al. (2008) reported a relative risk in males versus females of 1.4. Higher rates of schizophrenia are also seen among persons who are unmarried (Eaton et al. 1998). It has been hypothesized that this may be due to reverse causation, such that individuals at risk for schizophrenia are less likely to become married or more likely to divorce. The evidence is mixed; van Os et al. (2000) found that the incidence of schizophrenia was higher in single people than married people.

There is also strong evidence for a role of perinatal risk factors in the etiology of schizophrenia. In a meta-analysis of 18 studies, the risk for schizophrenia approximately doubled (odds ratio of 2) for infants born with obstetric complications of all kinds (Geddes and Lawrie 1995). Susser and Lin (1992) showed that nutrition deprivation during pregnancy was associated with increased risk of schizophrenia in offspring. Other risk factors include being born in the winter (Torrey et al. 1993), urbanicity (McGrath et al. 2004), and older paternal age (Byrne et al. 2003). With regard to course and prognosis, many studies find that younger patients, males, and those with a family history have worse outcomes (Tsuang and Tohen 2002).

Current epidemiological investigations regarding the etiology of schizophrenia, as reflected by the work of two of the current authors, demonstrate the range of interdisciplinary approaches used in contemporary psychiatric epidemiology. A series of cohort studies dating back to a paper by Susser and Lin (1992), conducted first in the Netherlands (Susser et al. 2012) and later in China (St Clair et al. 2005) have identified periconceptional maternal nutritional deprivation (approximately 4 weeks before to 8 weeks after conception) as a risk factor for subsequent schizophrenia. Susser and colleagues are pursuing a program of epigenetic research linking maternal prenatal nutrition, DNA methylation and schizophrenia (Kirkbride et al. 2012). Several studies in humans now suggest that prenatal nutritional exposures can influence DNA methylation, in offspring, although the evidence is considerably more scant than that provided by animal studies. A follow-up study of persons exposed and unexposed to famine during the Dutch Hunger Winter of 1944–1945 (not focused on schizophrenia) reported epigenetic changes among the offspring of women exposed to periconceptional starvation. The “sensitive” period for these epigenetic changes was very similar to the “sensitive” period identified in the schizophrenia studies described above (Heijmans et al. 2008). At approximately 60 years old, these offspring exhibited less methylation of the IGF2 locus in whole-blood samples than offspring of women who were unexposed or exposed later in gestation. IGF2 is an imprinted gene that plays a crucial role in early development and continues to play a role in cognitive and other brain processes over the life course (Chen et al. 2011). If genuine, it suggests that periconceptional maternal famine exposure can have a lasting effect on offspring DNA methylation over the life course. More recent analysis of global methylation patterns in this cohort have demonstrated no discernible difference in global DNA methylation patterns between offspring of exposed and unexposed mothers (Lumey et al. 2012), perhaps highlighting the complexity and potentially target-specific nature of prenatal exposures upon the fetal epigenome. A further important question now being examined is whether or not the effects of prenatal nutritional deficiency are transmitted across generations (Susser et al. 2012).

Buka and colleagues (1999) are also engaged in a wide array of epidemiological studies investigating the direct and interactive effects of conditions during pregnancy and familial and genetic risks in the etiology of schizophrenia. These involve a number of study designs, high-dimensional measurement, and statistical approaches. Study designs include basic cohort, nested case-control, sibling, high-risk, and multistage approaches where participants receive greater or lesser intensity of assessment based upon their risk profiles. Study measures reflect the influence of previous generations of epidemiological research (e.g., sociodemographic factors) as well as influences from related disciplines including neuropsychology, neuroscience, genetics, and immunology. Assessments include traditional diagnostic procedures, but also more fine-grained assessments of presumed related endophenotypes, including neuropsychological testing, functional and structural brain imaging, molecular genetics, and immune function. Following decades of research suggesting prenatal risks for schizophrenia, these and other investigators have used archived maternal serum samples from long-running pregnancy cohort studies to confirm the association between maternal infections, exposure to inflammatory cytokines during pregnancy, and subsequent schizophrenia (Buka et al. 2001a, b; Brown et al. 2004a, b). Subsequent work has incorporated molecular genetic methods, investigating the combined influences of prenatal complications, prenatal neurotrophic factors which are stimulated as part of a neuroprotective response to fetal distress (e.g., brain-derived neurotrophic factor), and related polymorphisms which have been associated with schizophrenia (Cannon et al. 2008). Other work motivated by this combined neurodevelopmental, neuroscience, and epidemiological orientation has focused on established gender differences in the phenomenology of schizophrenia and fetal brain development (Goldstein et al. 2011). New analyses move beyond traditional diagnostic classifications of psychotic disorders to dimensional approaches characterized by level and severity of dysfunction in multiple domains (Goldstein et al. 2001; Seidman et al. 2006, 2012). This is in line with recent developments at the National Institute of Mental Health calling for new Research Domain Criteria (RDoC) with efforts to integrate genetic, neurobiological, imaging, behavioral, and clinical data. Modern epidemiological approaches require such integration of new disciplines, designs, and analyses to continue to advance understanding of the epidemiology of schizophrenia.

Prevention Although preventive interventions are beyond the scope of this chapter, we believe it important to note that there are already examples of successful public health prevention efforts for mental disorders, and promising initiatives that may lead to discovery of others. To illustrate, we use the example of nutritional interventions.

We offer two examples of nutritional interventions already implemented and proven effective. Pellagra was once a major cause of mental disorder, especially psychosis (Susser et al. 2006). After public health programs addressed the micronutrient deficiency underlying pellagra, it became a rare cause, almost nonexistent in high-income countries. Similarly, and equally dramatic, prenatal iodine deficiency was shown to be a cause of intellectual disability, mild cognitive impairment, and behavioral problems (Zimmermann 2012). Public health interventions such as iodized salt have dramatically reduced mental impairments related to iodine deficiency, in most though unfortunately still not all parts of the globe.

With regard to promising initiatives, we offer an example that follows on the studies linking periconceptional starvation and schizophrenia described above. Since neural tube defects are related to periconceptional folic acid intake and peaked in the same birth cohort as schizophrenia in the Dutch Famine studies, it is natural to consider whether folic acid deficiency also played a role in schizophrenia. This hypothesis is being explored in a variety of ways, one of which is the ongoing follow-up of a large Norwegian cohort comprising more than 100,000 pregnancies (Magnus et al. 2006; Stoltenberg et al. 2010). Studies from the Norwegian cohort have found that maternal use of periconceptional folic acid supplements (from 4 weeks before to 8 weeks after the last menstrual period) is associated with a reduced risk of child neurodevelopmental disorders such as autism and severe language delay (Surén et al. 2013). Moreover, as noted earlier, converging lines of evidence suggest substantial overlap in the causes of autism and schizophrenia. Within the next decade, the ongoing follow-up of this Norwegian cohort will make it possible to directly test whether these supplements are related to reduced risk of schizophrenia and, if so, to examine potential mechanisms (Kirkbride et al. 2012). We only consider this a promising initiative because we do not yet know whether periconceptional folic acid supplements will ultimately be proven to reduce the risk of autism, schizophrenia, and/or other neurodevelopmental disorders after birth (they are already proven to reduce the risk of neural tube defects at birth). Thus far, however, the evidence for this hypothesis is promising, and if it is ultimately proven true, it will add substantially to examples of successful prevention efforts achieved through public health initiatives.

59.5 Conclusions

Based on generations of studies, it is quite clear that the burden of mental disorders is very high across the globe, in LMIC as well as in high-income countries. Indeed, for several reasons, the burden is likely to increase substantially in the coming decades. For example, as we continue recent progress in reducing child mortality in LMIC, a large population of "child survivors" emerges, who often have neurodevelopmental delays and disabilities (Scherzer et al. 2012). Yet the resources devoted to research and treatment of mental disorders remains minimal in comparison with the resources devoted to other disorders, which make a smaller contribution to burden of disease (Prince et al. 2007; Saxena et al. 2007). The reasons for this discrepancy are not fully understood, but surely one is the stigma associated with mental illness. Since the shortage of resources is most pronounced in LMIC, a current focus of WHO, NIMH, and other leading organizations is to identify strategies to "close the treatment gap" in LMIC (Collins et al. 2011). A related problem is that services that do exist are not well distributed within countries; at times, those with no diagnosis of a mental disorder are more likely to access psychiatric treatment than those with severe mental illness.

We propose that we are now at a critical moment of transformation in psychiatric epidemiology. In recent years, epidemiological investigations of the causes of mental disorders have grown exponentially, and the nature of these investigations has changed. We are now studying specific genetic and epigenetic causes and their relationship to environmental exposures, using large population registries that provide data over long time periods, employing new technologies to examine neighborhood and societal-level effects, and transforming the use of high-risk and other designs that previously existed but could not be applied with the same sophistication. This kind of research entails interdisciplinary collaboration (and training to facilitate this), refined assessment of potential etiological factors, and more attention to potential mechanisms. It also entails the acknowledgement that factors at multiple levels, from cell to society, contribute to these disorders. This broader framework is important for the design of all studies; even though a single study will likely address questions at only one or two levels, the framework helps in the selection of the appropriate level and the understanding of the limitations inherent in studying only one level (Susser and Susser 1996; Susser et al. 2006).

The recognition that mental disorders should be a priority in all populations has also grown exponentially. Although the United Nations in 2011 did not include mental disorders among non-communicable disease priorities, there was considerable controversy about this exclusion, and soon afterward many international funding agencies added mental disorders to their list of priorities. The 2011 decision by the UN is already being seen as a mistake by many leaders in public health. We think it highly unlikely that by the end of this decade, the exclusion of mental disorders will still be considered acceptable in any broad public health initiative.

The magnitude of the global public health burden associated with mental disorders also highlights the need for additional applied research related to the organization, financing, and delivery of mental health services. Throughout the past 100 years of research, there has been consistent evidence of the sizable proportion of persons with diagnosable mental disorders who receive no or limited treatment. At the same time, more recent epidemiological surveys reflect a high level of "met unneed" where individuals with low levels of need had a higher probability of accessing treatment than those with more severe conditions (Kessler et al. 1997; Mojtabai et al. 2002). There is an urgent need to increase both research and mental health services throughout the continuum of prevention and treatment, to increase availability in LMIC, and to better align service receipt with treatment need. As highlighted in this chapter, the field of epidemiology has advanced our knowledge of the forms and prevalence of major psychiatric conditions and is contributing to new discoveries regarding etiology. Parallel emphasis is needed on the applied side of psychiatric epidemiology, around the distribution and evaluation of population-level resources to reduce the burden of these highly prevalent and debilitating set of disorders.

References

Afifi TO, Cox BJ, Enns MW (2006) Mental health profiles among married, never-married, and seperated/divorced mothers in a nationally representative sample. Soc Psychiatry Psychiatr Epidemiol 41(2):122–129

Almeida-Filho N, Mari Jde J, Coutinho E, Franca JF, Fernandes J, Andreoli SB, Busnello ED (1997) Brazilian multicentric study of psychiatric morbidity. Methodological features and prevalence estimates. Br J Psychiatry 171:524–529

Alonso J, Angermeyer MC, Bernert S, Bruffaerts R, Brugha TS, Bryson H, de Girolamo G, Graaf R, Demyttenaere K, Gasquet I, Haro JM, Katz SJ, Kessler RC, Kovess V, Lépine JP, Ormel J, Polidori G, Russo LJ, Vilagut G, Almansa J, Arbabzadeh-Bouchez S, Autonell J, Bernal M, Buist-Bouwman MA, Codony M, Domingo-Salvany A, Ferrer M, Joo SS, Martínez-Alonso M, Matschinger H, Mazzi F, Morgan Z, Morosini P, Palacín C, Romera B, Taub N, Vollebergh WA; ESEMeD/MHEDEA 2000 Investigators, European Study of the Epidemiology of Mental Disorders (ESEMeD) Project (2004) Disability and quality of life impact of mental disorders in Europe: results from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project. Acta Psychiatr Scand Suppl (420):38–46

Andrews G, Hall W, Teesson M, Henderson S (1999) The mental health of Australians. Australian Bureau of Statistics, Canberra

Angold A, Costello EJ (1995) Developmental epidemiology. Epidemiol Rev 17(1):74–82

Bailey A, Le Couteur A, Gottesman I, Bolton P, Simonoff E, Yuzda E, Rutter M (1995) Autism as a strongly genetic disorder: evidence from a British twin study. Psychol Med 25(1):63–77

Baron-Cohen S, Scott FJ, Allison C, Williams J, Bolton P, Matthews FE, Brayne C (2009) Prevalence of autism-spectrum conditions: UK school-based population study. Br J Psychiatry 194(6):500–509

Bjorkenstam E, Hallgvist J, Dalman C, Ljung R (2012) Risk of new psychiatric episodes in the year following divorce in midlife: cause or selection? A nationwide register-based study of 703,960 individuals. Int J Soc Psychiatry (Epub ahead of print)

Bresnahan M, Li G, Susser E (2009) Hidden in plain sight. Int J Epidemiol 38(5):1172–1174

Breton JJ, Bergeron L, Valla JP, Berthiaume C, Gaudet N, Lambert J, St-Georges M, Houde L, Lépine S (1999) Quebec child mental health survey: prevalence of DSM-III-R mental health disorders. J Child Psychol Psychiatry 40(3):375–384

Bromet EJ, Schwartz JE, Fennig S, Geller L, Jandorf L, Kovasznay B, Lavelle J, Miller A, Pato C, Ram R, Rich C (1992) The epidemiology of psychosis: the Suffolk County Mental Health Project. Schizophr Bull 18(2):243–255

Brown AS, Begg MD, Gravenstein S, Schaefer CA, Wyatt RJ, Bresnahan M, Babulas VP, Susser ES (2004a) Serologic evidence of prenatal influenza in the etiology of schizophrenia. Arch Gen Psychiatry 61(8):774–780

Brown AS, Hooton J, Schaefer CA, Zhang H, Petkova E, Babulas V, Perrin M, Gorman JM, Susser ES (2004b) Elevated maternal interleukin-8 levels and risk of schizophrenia in adult offspring. Am J Psychiatry 161(5):889–895

Buka SL, Tsuang MT, Lipsitt LP (1993) Pregnancy/delivery complications and psychiatric diagnosis. A prospective study. Arch Gen Psychiatry 50(2):151–156

Buka SL, Goldstein JM, Seidman LJ, Zornberg GL, Donatelli JA, Denny LR, Tsuang MT (1999) Prenatal complications, genetic vulnerability and schizophrenia: the New England longitudinal studies of schizophrenia. Psychiatr Ann 29(3):151–156

Buka SL, Tsuang MT, Torrey EF, Klebanoff MA, Bernstein D, Yolken RH (2001a) Maternal infections and subsequent psychosis among offspring. Arch Gen Psychiatry 58(11):1032–1037

Buka SL, Tsuang MT, Torrey EF, Klebanoff MA, Wagner RL, Yolken RH (2001b) Maternal cytokine levels during pregnancy and adult psychosis. Brain Behav Immun 15(4):411–420

Burns BJ, Costello EJ, Angold A, Tweed D, Stangl D, Farmer E M, Erkanli A (1995) Children's mental health service use across service sectors. Health Aff 14(3):147–159

Byrne M, Agerbo E, Ewald H, Eaton WW, Mortensen PB (2003) Parental age and risk of schizophrenia: a case-control study. Arch Gen Psychiatry 60(7):673–678

Canino G, Shrout PE, Rubio-Stipec M, Bird HR, Bravo M, Ramirez R, Chavez L, Alegria M, Bauermeister JJ, Hohmann A, Ribera J, Garcia P, Martinez-Taboas A (2004) The DSM-IV rates of child and adolescent disorders in Puerto Rico: prevalence, correlates, service use, and the effects of impairment. Arch Gen Psychiatry 61(1):85–93

Cannon TD, Kaprio J, Lonnqvist J, Huttunen M, Koskenvuo M (1998) The genetic epidemiology of schizophrenia in a Finnish twin cohort. A population-based modeling study. Arch Gen Psychiatry 55(1):67–74

Cannon TD, Yolken R, Buka S, Torrey EF (2008) Decreased neurotrophic response to birth hypoxia in the etiology of schizophrenia. Biol Psychiatry 64(9):797–802

Chakrabarti S, Fombonne E (2001) Pervasive developmental disorders in preschool children. JAMA 285(24):3093–3099

Charman T, Pickles A, Chandler S, Wing L, Bryson S, Simonoff E, Loucas T, Baird G (2009) Commentary: effects of diagnostic thresholds and research vs service and administrative diagnosis on autism prevalence. Int J Epidemiol 38(5):1234–1238; author reply 1243–1244

Chen DY, Stern SA, Garcia-Osta A, Saunier-Rebori B, Pollonini G, Bambah-Mukku D, Blitzer RD, Alberini CM (2011) A critical role for IGF-II in memory consolidation and enhancement. Nature 469(7331):491–497

Chudley AE, Gutierrez E, Jocelyn LJ, Chodirker BN (1998) Outcomes of genetic evaluation in children with pervasive developmental disorder. J Dev Behav Pediatr 19(5):321–325

Cohen P, Cohen J, Kasen S, Velez CN, Hartmark C, Johnson J, Rojas M, Brook J, Streuning EL (1993) An epidemiological study of disorders in late childhood and adolescence – I. Age- and gender-specific prevalence. J Child Psychol Psychiatry 34(6):851–867

Collins PY, Patel V, Joestl SS, March D, Insel TR, Daar AS; Scientific Advisory Board and the Executive Committee of the Grand Challenges on Global Mental Health, Anderson W, Dhansay MA, Phillips A, Shurin S, Walport M, Ewart W, Savill SJ, Bordin IA, Costello EJ, Durkin M, Fairburn C, Glass RI, Hall W, Huang Y, Hyman SE, Jamison K, Kaaya S, Kapur S, Kleinman A, Ogunniyi A, Otero-Ojeda A, Poo MM, Ravindranath V, Sahakian BJ, Saxena S, Singer PA, Stein DJ (2011) Grand challenges in global mental health. Nature 475(7354):27–30

Corcoran C, Perrin M, Harlap S, Deutsch L, Fennig S, Manor O, Nahon D, Kimhy D, Malaspina D, Susser E (2009) Effect of socioeconomic status and parents' education at birth on risk of schizophrenia in offspring. Soc Psychiatry Psychiatr Epidemiol 44(4):265–271

Costello EJ, Angold A, Burns BJ, Erkanli A, Stangl DK, Tweed DL (1996) The Great Smoky Mountains Study of Youth. Functional impairment and serious emotional disturbance. Arch Gen Psychiatry 53(12):1137–1143

Costello EJ, Mustillo S, Erkanli A, Keeler G, Angold A (2003) Prevalence and development of psychiatric disorders in childhood and adolescence. Arch Gen Psychiatry 60(8):837–844

Costello EJ, Mustillo S, Keller G, Angold A (2004) Prevalence of psychiatric disorders in childhood and adolescence. In: Levin BL, Petrila J, Hennessy KD (eds) Mental health services: a public health perspective, 2nd edn. Oxford University Press, Oxford, pp 111–128

Costello EJ, Egger H, Angold A (2005) 10-year research update review: the epidemiology of child and adolescent psychiatric disorders: I. Methods and public health burden. J Am Acad Child Adolesc Psychiatry 44(10):972–986

Croen LA, Grether JK, Selvin S (2002) Descriptive epidemiology of autism in a California population: who is at risk? J Autism Dev Disord 32(3):217–224

Danielsson S, Gillberg IC, Billstedt E, Gillberg C, Olsson I (2005) Epilepsy in young adults with autism: a prospective population-based follow-up study of 120 individuals diagnosed in childhood. Epilepsia 46(6):918–923

Dawson G, Rogers S, Munson J, Smith M, Winter J, Greenson J, Donaldson A, Varley J (2010) Randomized, controlled trial of an intervention for toddlers with autism: the Early Start Denver Model. Pediatrics 125(1):e17–23

Demyttenaere K, Bruffaerts R, Posada-Villa J, Gasquet I, Kovess V, Lepine JP, Angermeyer MC, Bernert S, de Girolamo G, Morosini P, Polidori G, Kikkawa T, Kawakami N, Ono Y, Takeshima T, Uda H, Karam EG, Fayyad JA, Karam AN, Mneimneh ZN, Medina-Mora ME, Borges G, Lara C, de Graaf R, Ormel J, Gureje O, Shen Y, Huang Y, Zhang M, Alonso J, Haro JM, Vilagut G, Bromet EJ, Gluzman S, Webb C, Kessler RC, Merikangas KR, Anthony JC, Von Korff MR, Wang PS, Brugha TS, Aguilar-Gaxiola S, Lee S, Heeringa S, Pennell BE, Zaslavsky AM, Ustun TB, Chatterji S; WHO World Mental Health Survey Consortium (2004) Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. JAMA 291(21):2581–2590

Dohrenwend BP (1997) A psychosocial perspective on the past and future of psychiatric epidemiology. Am J Epidemiol 147:222–231

Dohrenwend BP, Dohrenwend BS (1982) Perspectives on the past and future of psychiatric epidemiology. The 1981 Rema Lapouse Lecture. Am J Public Health 72(11):1271–1279

Dohrenwend BP, Levav I, Shrout PE, Schwartz S, Naveh G, Link BG, Skodol AE, Stueve A (1992) Socioeconomic status and psychiatric disorders: the causation-selection issue. Science 255(5047):946–952

Duchan E, Patel DR (2012) Epidemiology of autism spectrum disorders. Pediatr Clin North Am 59(1):27–43, ix–x

Durkheim E (1951) Suicide: a study in sociology. Free Press, Glencoe

Durkin MS, Maenner MJ, Newschaffer CJ, Lee LC, Cunniff CM, Daniels JL, Kirby RS, Leavitt L, Miller L, Zahorodny W, Schieve LA (2008) Advanced parental age and the risk of autism spectrum disorder. Am J Epidemiol 168(11):1268–1276

Eaton W (ed) (2012) Public mental health, vol 1. Oxford University Press, New York

Eaton WW, Day R, Kramer M (1998) The use of epidemiology for risk factor research in schizophrenia: an overview and methodologic critique. In: Tsuang M, Simpson JC (eds) Handbook of schizophrenia, vol 3. Elsevier Science, Amsterdam, pp 169–204

Faris RELDH (1939) Mental disorders in urban areas: an ecological study of schizophrenia and other psychoses. University of Chicago Press, Chicago

Fombonne E (2009) Epidemiology of pervasive developmental disorders. Pediatr Res 65(6): 591–598

Fox JW (1980) Gove's specific sex-role theory of mental illness: a research note. J Health Soc Behav 21:260–266

Gardener H, Spiegelman D, Buka SL (2009) Prenatal risk factors for autism: comprehensive meta-analysis. Br J Psychiatry 195:7–14

Geddes JR, Lawrie SM (1995) Obstetric complications and schizophrenia: a meta-analysis. Br J Psychiatry 167(6):786–793

Gillberg C (1984) Infantile autism and other childhood psychoses in a Swedish urban region. Epidemiological aspects. J Child Psychol Psychiatry 25(1):35–43

Gillberg C, Billstedt E, Sundh V, Gillberg IC (2010) Mortality in autism: a prospective longitudinal community-based study. J Autism Dev Disord 40(3):352–357

Gilman SR, Chang J, Xu B, Bawa TS, Gogos JA, Karayiorgou M, Vitkup (2012) Diverse types of genetic variation converge on functional gene networks involved in schizophrenia. Nat Neurosci 15:1723–1728

Goldstein JM, Seidman LJ, Horton NJ, Makris N, Kennedy DN, Caviness VS Jr, Faraone SV, Tsuang MT (2001) Normal sexual dimorphism of the adult human brain assessed by in vivo magnetic resonance imaging. Cereb Cortex 11(6):490–497

Goldstein JM, Cherkerzian S, Seidman LJ, Petryshen TL, Fitzmaurice G, Tsuang MT, Buka SL (2011) Sex-specific rates of transmission of psychosis in the New England high-risk family study. Schizophr Res 128(1–3):150–155

Gottesman II, Shields J (1982) Schizophrenia: the epigenetic puzzle. Cambridge University Press, Cambridge

Grant BF, Harford TC (1995) Comorbidity between DSM-IV alcohol use disorders and major depression: results of a national survey. Drug Alcohol Depend 39(3):197–206

Grant BF, Stinson FS, Dawson DA, Chou SP, Dufour MC, Compton W, Pickering RP, Kaplan K (2004) Prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders: results from the National Epidemiologic Survey on Alcohol and Related Conditions. Arch Gen Psychiatry 61(8):807–816

Grob GN (2005) Public policy and mental illnesses: Jimmy Carter's Presidential Commission on Mental Health. Milbank Q 83(3):425–456

Hallmayer J, Cleveland S, Torres A, Phillips J, Cohen B, Torigoe T, Miller J, Fedele A, Collins J, Smith K, Lotspeich L, Croen LA, Ozonoff S, Lajonchere C, Grether JK, Risch N (2011) Genetic heritability and shared environmental factors among twin pairs with autism. Arch Gen Psychiatry 68(11):1095–1102

Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, Slagboom PE, Lumey LH (2008) Persistent epigenetic differences associated with prenatal exposure to famine in humans. Proc Natl Acad Sci U S A 105(44):17046–17049

Helzer JE, Robins LN, McEvoy LT, Spitznagel EL, Stoltzman RK, Farmer A, Brockington IF (1985) A comparison of clinical and diagnostic interview schedule diagnoses. Physician reexamination of lay-interviewed cases in the general population. Arch Gen Psychiatry 42(7):657–666

Heston LL (1966) Psychiatric disorders in foster home reared children of schizophrenic mothers. Br J Psychiatry 112(489):819–825

Heussler H, Polnay L, Marder E, Standen P, Chin LU, Butler N (2001) Prevalence of autism in early 1970s may have been underestimated. BMJ 323(7313):633

Hollingshead AB, Redlich FC (1958) Social class and mental illness. Wiley, New York

Honda H, Shimizu Y, Imai M, Nitto Y (2005) Cumulative incidence of childhood autism: a total population study of better accuracy and precision. Dev Med Child Neurol 47(1):10–18

Hopper K, Wanderling J (2000) Revisiting the developed versus developing country distinction in course and outcome in schizophrenia: results from ISoS, the WHO Collaborative Followup Project. Schizophr Bull 26(4):835–846

Hultman CM, Sparen P, Cnattingius S (2002) Perinatal risk factors for infantile autism. Epidemiology 13(4):417–423

Hwu HG, Yeh EK, Chang LY (1989) Prevalence of psychiatric disorders in Taiwan defined by the Chinese Diagnostic Interview Schedule. Acta Psychiatr Scand 79(2):136–147

Insel TR, Fenton WS (2005) Psychiatric epidemiology: it's not just about counting anymore. Arch Gen Psychiatry 62(6):590–592

Jablensky A, Sartorius N, Ernberg G, Anker M, Korten A, Cooper J, Day R, Bertelsen A (1992) Schizophrenia: manifestations, incidence and course in different cultures. A World Health Organization ten-country study. Psychol Med Monogr Suppl 20:1–97

Jarvis E (1971) Insanity and idiocy in Massachusetts: report of the Commission of Lunacy, 1855. Harvard University Press, Cambridge

Jones P, Cannon M (1998) The new epidemiology of schizophrenia. Psychiatr Clin North Am 21(1):1–25

Kawamura Y, Takahashi O, Ishii T (2008) Reevaluating the incidence of pervasive developmental disorders: impact of elevated rates of detection through implementation of an integrated system of screening in Toyota, Japan. Psychiatry Clin Neurosci 62(2):152–159

Kebede D, Alem A, Shibre T, Negash A, Fekadu A, Fekadu D, Deyassa N, Jacobsson L, Kullgren G (2003) Onset and clinical course of schizophrenia in Butajira-Ethiopia – a community-based study. Soc Psychiatry Psychiatr Epidemiol 38(11):625–631

Keller MB, Lavori PW, Mueller TI, Endicott J, Coryell W, Hirschfeld RM, Shea T (1992) Time to recovery, chronicity, and levels of psychopathology in major depression. A 5-year prospective follow-up of 431 subjects. Arch Gen Psychiatry 49(10):809–816

Keller MB, Lavori PW, Coryell W, Endicott J, Mueller TI (1993) Bipolar I: a five-year prospective follow-up. J Nerv Ment Dis 181(4):238–245

Kendler KS, Davis CG, Kessler RC (1997) The familial aggregation of common psychiatric and substance use disorders in the National Comorbidity Survey: a family history study. Br J Psychiatry 170:541–548

Kessler RC (1994) Building on the ECA: the National Comorbidity Survey and the Children's ECA. Int J Methods Psychiatr Res 4(2):81–94

Kessler RC, Walters EE (1998) Epidemiology of DSM-III-R major depression and minor depression among adolescents and young adults in the National Comorbidity Survey. Depress Anxiety 7(1):3–14

Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, Wittchen HU, Kendler KS (1994) Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: results from the National Comorbidity Survey. Arch Gen Psychiatry 51(1):8–19

Kessler RC, Frank RG, Edlund M, Katz SJ, Lin E, Leaf P (1997) Differences in the use of psychiatric outpatient services between the United States and Ontario. N Engl J Med 336(8):551–557

Kessler RC, Berglund PA, Bruce ML, Koch JR, Laska EM, Leaf PJ, Manderscheid RW, Rosenheck RA, Walters EE, Wang PS (2001) The prevalence and correlates of untreated serious mental illness. Health Serv Res 36(6 Pt 1):987–1007

Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE (2005a) Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 62(6):593–602

Kessler RC, Chiu WT, Demler O, Walters EE (2005b) Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 62(6):617–627

Kessler RC, Chiu WT, Jin R, Ruscio AM, Shear K, Walters EE (2006) The epidemiology of panic attacks, panic disorder, and agoraphobia in the National Comorbidity Survey Replication. Arch Gen Psychiatry 63(4):415–424

Kessler RC, Amminger GP, Aguilar-Gaxiola S, Alonso J, Lee S, Ustun TB (2007) Age of onset of mental disorders: a review of recent literature. Curr Opin Psychiatry 20(4):359–364

Kessler RC, Heeringa S, Lakoma MD, Petukhova M, Rupp AE, Schoenbaum M, Wang PS, Zaslavsky AM (2008) Individual and societal effects of mental disorders on earnings in the United States: results from the national comorbidity survey replication. Am J Psychiatry 165(6):703–711

Kessler RC, Avenevoli S, Costello EJ, Georgiades K, Green JG, Gruber MJ, He JP, Koretz D, McLaughlin KA, Petukhova M, Sampson NA, Zaslavsky AM, Merikangas KR (2012) Prevalence, persistence, and sociodemographic correlates of DSM-IV disorders in the National Comorbidity Survey Replication Adolescent Supplement. Arch Gen Psychiatry 69(4):372–380

Keyes KM, Susser E, Cheslack-Postava K, Fountain C, Liu K, Bearman PS (2012) Cohort effects explain the increase in autism diagnosis among children born from 1992 to 2003 in California. Int J Epidemiol 41(2):495–503

Kim YS, Leventhal BL, Koh YJ, Fombonne E, Laska E, Lim EC, Cheon KA, Kim SJ, Kim YK, Lee H, Song DH, Grinker RR (2011) Prevalence of autism spectrum disorders in a total population sample. Am J Psychiatry 168(9):904–912

King M, Bearman P (2009) Diagnostic change and the increased prevalence of autism. Int J Epidemiol 38(5):1224–1234

Kirkbride JB, Susser E, Kundakovic M, Kresovich JK, Davey Smith G, Relton CL (2012) Prenatal nutrition, epigenetics and schizophrenia risk: can we test causal effects? Epigenomics 4(3): 303–315

Kogan MD, Blumberg SJ, Schieve LA, Boyle CA, Perrin JM, Ghandour RM, Singh GK, Strickland BB, Trevathan E, van Dyck PC (2009) Prevalence of parent-reported diagnosis of autism spectrum disorder among children in the US, 2007. Pediatrics 124(5):1395–1403

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, Wong WS, Sigurdsson G, Walters GB, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Magnusson OT, Thorsteinsdottir U, Stefansson K (2012) Rate of de novo mutations and the importance of father's age to disease risk. Nature 488(7412):471–475

Kumar RA, Christian SL (2009) Genetics of autism spectrum disorders. Curr Neurol Neurosci Rep 9(3):188–197

Kumar RA, KaraMohamed S, Sudi J, Conrad DF, Brune C, Badner JA, Gilliam TC, Nowak NJ, Cook EH Jr, Dobyns WB, Christian SL (2008) Recurrent 16p11.2 microdeletions in autism. Hum Mol Genet 17(4):628–638

Langner TS (1962) A twenty-two item screening score of psychiatric symptoms indicating impairment. J Health Hum Behav 3:269–276

Lauritsen MB, Pedersen CB, Mortensen PB (2005) Effects of familial risk factors and place of birth on the risk of autism: a nationwide register-based study. J Child Psychol Psychiatry 46(9): 963–971

Leighton DC, Harding JS, Macklin D, Hughes CC, Leighton AH (1963) Psychiatric findings of the Stirling County Study. Am J Psychiatry 119(11):1021–1026

Levy SE, Giarelli E, Lee LC, Schieve LA, Kirby RS, Cunniff C, Nicholas J, Reaven J, Rice CE (2010) Autism spectrum disorder and co-occurring developmental, psychiatric, and medical conditions among children in multiple populations of the United States. J Dev Behav Pediatr 31(4):267–275

Lichtenstein P, Carlstrom E, Rastam M, Gillberg C, Anckarsater H (2010) The genetics of autism spectrum disorders and related neuropsychiatric disorders in childhood. Am J Psychiatry 167(11):1357–1363

Lord C, Risi S, Lambrecht L, Cook EH Jr, Leventhal BL, DiLavore PC, Pickles A, Rutter M (2000) The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. J Autism Dev Disord 30(3):205–223

Lord C, Petkova E, Hus V, Gan W, Lu F, Martin DM, Ousley O, Guy L, Bernier R, Gerdts J, Algermissen M, Whitaker A, Sutcliffe JS, Warren Z, Klin A, Saulnier C, Hanson E, Hundley R, Piggot J, Fombonne E, Steiman M, Miles J, Kanne SM, Goin-Kochel RP, Peters SU, Cook EH, Guter S, Tjernagel J, Green-Snyder LA, Bishop S, Esler A, Gotham K, Luyster R, Miller F, Olson J, Richler J, Risi S (2012) A multisite study of the clinical diagnosis of different autism spectrum disorders. Arch Gen Psychiatry 69(3):306–313

Lumey LH, Terry MB, Delgado-Cruzata L, Liao Y, Wang Q, Susser E, McKeague I, Santella RM (2012) Adult global DNA methylation in relation to pre-natal nutrition. Int J Epidemiol 41(1):116–123

Magnus P, Irgens LM, Haug K, Nystad W, Skjaerven R, Stoltenberg C; MoBa Study Group (2006) Cohort profile: the Norwegian Mother and Child Cohort Study (MoBa). Int J Epidemiol 35(5):1146–1150

Manning SE, Davin CA, Barfield WD, Kotelchuck M, Clements K, Diop H, Osbahr T, Smith LA (2011) Early diagnoses of autism spectrum disorders in Massachusetts birth cohorts, 2001–2005. Pediatrics 127(6):1043–1051

Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, Shago M, Moessner R, Pinto D, Ren Y, Thiruvahindrapduram B, Fiebig A, Schreiber S, Friedman J, Ketelaars CE, Vos YJ, Ficicioglu C, Kirkpatrick S, Nicolson R, Sloman L, Summers A, Gibbons CA, Teebi A, Chitayat D, Weksberg R, Thompson A, Vardy C, Crosbie V, Luscombe S, Baatjes R, Zwaigenbaum L, Roberts W, Fernandez B, Szatmari P, Scherer SW (2008) Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet 82(2):477–488

McClellan JM, Susser E, King MC (2007) Schizophrenia: a common disease caused by multiple rare alleles. Br J Psychiatry 190:194–199

McGrath J, Saha S, Welham J, El Saadi O, MacCauley C, Chant D (2004) A systematic review of the incidence of schizophrenia: the distribution of rates and the influence of sex, urbanicity, migrant status and methodology. BMC Med 2:13

McGrath J, Saha S, Chant D, Welham J (2008) Schizophrenia: a concise overview of incidence, prevalence, and mortality. Epidemiol Rev 30:67–76

McPartland JC, Reichow B, Volkmar FR (2012) Sensitivity and specificity of proposed DSM-5 diagnostic criteria for autism spectrum disorder. J Am Acad Child Adolesc Psychiatry 51(4):368–383

Mehta P, Joseph A, Verghese A (1985) An epidemiologic study of psychiatric disorders in a rural area in tamilnadu. Ind J Psychiatry 27(2):153–158

Menezes PR, Scazufca M, Busatto GF, Coutinho LMS, McGuire PK, Murray RM (2007) Incidence of first-contact psychosis in Sao Paulo, Brazil. Br J Psychiatry 191(Suppl 51):102–106

Merikangas KR, Avenevoli S (2002) Epidemiology of mood and anxiety disorders in children and adolescents. In: Tsaung M, Tohen M (eds) Textbook in psychiatric epidemiology, 2nd edn. Wiley-Liss, New York, pp 657–704

Merikangas KR, Nakamura EF, Kessler RC (2009a) Epidemiology of mental disorders in children and adolescents. Dialogues Clin Neurosci 11(1):7–20

Merikangas KR, Avenevoli S, Costello J, Koretz D, Kessler RC (2009b) National comorbidity survey replication adolescent supplement (NCS-A): I. Background and measures. J Am Acad Child Adolesc Psychiatry 48(4):367–369

Merikangas KR, He JP, Burstein M, Swanson SA, Avenevoli S, Cui L, Benjet C, Georgiades K, Swendsen J (2010) Lifetime prevalence of mental disorders in U.S. adolescents: results from the National Comorbidity Survey Replication–Adolescent Supplement (NCS-A). J Am Acad Child Adolesc Psychiatry 49(10):980–989

Miller JS, Bilder D, Farley M, Coon H, Pinborough-Zimmerman J, Jenson W, Rice CE, Fombonne E, Pingree CB, Ritvo E, Ritvo RA, McMahon WM (2013) Autism spectrum disorder reclassified: a second look at the 1980s Utah/UCLA Autism Epidemiologic Study. J Autism Dev Disord 43(1):200–210

Moffitt TE, Caspi A, Taylor A, Kokaua J, Milne BJ, Polanczyk G, Poulton R (2010) How common are common mental disorders? Evidence that lifetime prevalence rates are doubled by prospective versus retrospective ascertainment. Psychol Med 40(6):899–909

Mojtabai R, Olfson M, Mechanic D (2002) Perceived need and help-seeking in adults with mood, anxiety, or substance use disorders. Arch Gen Psychiatry 59(1):77–84

Morgan C, Fearon P (2007) Social experience and psychosis insights from studies of migrant and ethnic minority groups. Epidemiol Psichiatr Soc 16(2):118–123

Morrow EM (2010) Genomic copy number variation in disorders of cognitive development. J Am Acad Child Adolesc Psychiatry 49(11):1091–1104

Murray CJ, Lopez AD (1996) Evidence-based health policy – lessons from the Global Burden of Disease Study. Science 274(5288):740–743

Newman DL, Moffitt TE, Caspi A, Magdol L, Silva PA, Stanton WR (1996) Psychiatric disorder in a birth cohort of young adults: prevalence, comorbidity, clinical significance, and new case incidence from ages 11 to 21. J Consult Clin Psychol 64(3):552–562

Newschaffer C, Croen L, Fallin MD, Hertz-Picciotto I, Nguyen DV, Lee NL, Berry CA, Farzadegan H, Hess HN, Landa RJ, Levy SE, Massolo ML, Meyerer SC, Mohammed SM, Oliver MC, Ozonoff S, Pandey J, Schroeder A, Shedd-Wise KM (2012). Infant siblings and the investigation of autism risk factors. J Neurodev Disord 4(1):7

Oakley-Browne MA, Joyce PR, Wells JE, Bushnell JA, Hornblow AR (1989) Christchurch Psychiatric Epidemiology Study, part II: six month and other period prevalences of specific psychiatric disorders. Aust N Z J Psychiatry 23(3):327–340

Offord DR, Boyle MH, Szatmari P, Rae-Grant NI, Links PS, Cadman DT, Byles JA, Crawford JW, Blum HM, Byrne C, Thomas H, Woodward CA (1987) Ontario Child Health Study. II. Six-month prevalence of disorder and rates of service utilization. Arch Gen Psychiatry 44(9): 832–836

Owen MJ, Williams HJ, O'Donovan MC (2009) Schizophrenia genetics: advancing on two fronts. Curr Opin Genet Dev 19(3):266–270

Phillips MR, Yang G, Li S, Li Y (2004) Suicide and the unique prevalence pattern of schizophrenia in mainland China: a retrospective observational study. Lancet 364(9439):1062–1068

Pickett J, Xiu E, Tuchman R, Dawson G, Lajonchere C (2011) Mortality in individuals with autism, with and without epilepsy. J Child Neurol 26(8):932–939

Principal Investigators; Centers for Disease Control and Prevention (CDC) (2012) Prevalence of autism spectrum disorders – autism and developmental disabilities monitoring network, 14 sites, United States, 2008. MMWR Surveill Summ 61(3):1–19

Prince M, Patel V, Saxena S, Maj M, Maselko J, Phillips MR, Rahman A (2007) No health without mental health. Lancet 370(9590):859–877

Regier DA, Goldberg ID, Taube CA (1978) The de facto US Mental Health Services System. Arch Gen Psychiatry 35:685–693

Regier DA, Narrow WE, Rae DS, Manderscheid RW, Locke BZ, Goodwin FK (1993a) The de facto US mental and addictive disorders service system: epidemiologic catchment area prospective 1-year prevalence rates of disorders and services. Arch Gen Psychiatry 50(2):85–94

Regier DA, Farmer ME, Rae DS, Myers JK, Kramer M, Robins LN, George LK, Karno M, Locke BZ (1993b) One-month prevalence of mental disorders in the United States and sociodemographic characteristics: the Epidemiologic Catchment Area study. Acta Psychiatr Scand 88(1):35–47

Rivet TT, Matson JT (2011) Review of gender differences in core symptomatology in autism spectrum disorders. Res Autism Spectr Disord 5:957–976

Roberts RE, Attkisson CC, Rosenblatt A (1998) Prevalence of psychopathology among children and adolescents. Am J Psychiatry 155(6):715–725

Robins LN (1978) Psychiatric epidemiology. Arch Gen Psychiatry 35:697–702

Robins LN, Regier DA (eds) (1991) Psychiatric disorders in America: the Epidemiologic Catchment Area Study. The Free Press, New York

Robins LN, Helzer JE, Croughan J, Ratcliff KS (1981) National Institute of Mental Health Diagnostic Interview Schedule. Its history, characteristics, and validity. Arch Gen Psychiatry 38(4):381–389

Rosenberg RE, Law J, Yenokyan G, McGready J, Kaufrman WE, Law PA (2009) Characteristics and concordance of autism spectrum disorders among 277 twin pairs. Arch Pediatr Adolesc Med 163(10):907–914

Roth C, Magnus P, Schjolberg S, Stoltenberg C, Surén P, McKeague IW, Davey Smith G, Reichborn-Kjennerud T, Susser E (2011) Folic acid supplements in pregnancy and severe language delay in children. JAMA 306(14):1566–1573

Rutter M, Cox A, Tupling C, Berger M, Yule W (1975) Attainment and adjustment in two geographical areas. I – the prevalence of psychiatric disorder. Br J Psychiatry 126:493–509

Rutter M, Le Couteur A, Lord C (2003) ADI-R: autism diagnostic interview – revised: manual. Western Psychological Services, Los Angeles

Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, Walker MF, Ober GT, Teran NA, Song Y, El-Fishawy P, Murtha RC, Choi M, Overton JD, Bjornson RD, Carriero NJ, Meyer KA, Bilguvar K, Mane SM, Sestan N, Lifton RP, Günel M, Roeder K, Geschwind DH, Devlin B, State MW (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485(7397):237–241

Sartorius N, Shapiro R, Jablensky A (1974) The international pilot study of Schizophrenia. Schizophr Bull 1(11):21–34

Saxena S, Thornicroft G, Knapp M, Whiteford H (2007) Resources for mental health: scarcity, inequity, and inefficiency. Lancet 370(9590):878–889

Scherzer AL, Chhagan M, Kauchali S, Susser E (2012) Global perspective on early diagnosis and intervention for children with developmental delays and disabilities. Dev Med Child Neurol 54(12):1079–1084

Schmidt RJ, Hansen RL, Hartiala J, Allayee H, Schmidt LC, Tancredi DJ, Tassone F, Hertz-Picciotto I (2011) Prenatal vitamins, one-carbon metabolism gene variants, and risk for autism. Epidemiology 22(4):476–485

Schumann CM, Bloss CS, Barnes CC, Wideman GM, Carper RA, Akshoomoff N, Pierce K, Hagler D, Schork N, Lord C, Courchesne E (2010) Longitudinal magnetic resonance imaging study of cortical development through early childhood in autism. J Neurosci 30(12):4419–4427

Schwartz S, Susser E (2010) Genome-wide association studies: does only size matter? Am J Psychiatry 167:741–744

Seidman LJ, Buka SL, Goldstein JM, Tsuang MT (2006) Intellectual decline in schizophrenia: evidence from a prospective birth cohort 28 year follow-up study. J Clin Exp Neuropsychol 28(2):225–242

Seidman LJ, Cherkerzian S, Goldstein JM, Agnew-Blais J, Tsuang MT, Buka SL (2012) Neuropsychological performance and family history in children at age 7 who develop adult schizophrenia or bipolar psychosis in the New England Family Studies. Psychol Med 11:1–13

Seiler LH (1973) The 22-item scale used in field studies of mental illness: a question of method, a question of substance, and a question of theory. J Health Soc Behav 14(3):252–264

Shaffer D, Fisher P, Dulcan MK, Davies M, Piacentini J, Schwab-Stone ME, Lahey BB, Bourdon K, Jensen PS, Bird HR, Canino G, Regier DA (1996) The NIMH Diagnostic Interview Schedule for Children Version 2.3 (DISC-2.3): description, acceptability, prevalence rates, and performance in the MECA Study. J Am Acad Child Adolesc Psychiatry 35(7):865–877

Silva PA (1990) The Dunedin Multidisciplinary Health and Development Study: a 15 year longitudinal study. Paediatr Perinat Epidemiol 4(1):76–107

Smith T, Groen AD, Wynn JW (2000) Randomized trial of intensive early intervention for children with pervasive developmental disorder. Am J Ment Retard 105(4):269–285

Srole L, Langner TS, Michael ST, Opler ST, Rennie TAC (1962) Mental health in the metropolis: the Midtown Manhattan Study. McGraw-Hill, New York

St Clair D, Xu M, Wang P, Yu Y, Fang Y, Zhang F, Zheng X, Gu N, Feng G, Sham P, He L (2005) Rates of adult schizophrenia following prenatal exposure to the Chinese famine of 1959–1961. JAMA 294(5):557–562

State MW, Dykens EM (2000) Genetics of childhood disorders: XV. Prader-Willi syndrome: genes, brain, and behavior. J Am Acad Child Adolesc Psychiatry 39(6):797–800

Steffenburg S, Gillberg C, Hellgren L, Andersson L, Gillberg IC, Jakobsson G, Bohman M (1989) A twin study of autism in Denmark, Finland, Iceland, Norway and Sweden. J Child Psychol Psychiatry 30(3):405–416

Stoltenberg C, Schjolberg S, Bresnahan M, Hornig M, Hirtz D, Dahl C, Lie KK, Reichborn-Kjennerud T, Schreuder P, Alsaker E, Øyen AS, Magnus P, Surén P, Susser E, Lipkin WI; ABC Study Group (2010) The Autism Birth Cohort: a paradigm for gene-environment-timing research. Mol Psychiatry 15(7):676–680

Sullivan PF, Kendler KS, Neale MC (2003) Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Arch Gen Psychiatry 60(12):1187–1192

Sullivan PF, Magnusson C, Reichenberg A, Boman M, Dalman C, Davidson M, Fruchter E, Hultman CM, Lundberg M, Långström N, Weiser M, Svensson AC, Lichtenstein P (2012) Family history of schizophrenia and bipolar disorder as risk factors for AutismFamily history of psychosis as risk factor for ASD. Arch Gen Psychiatry Jul 2:1–5

Surén P, Roth C, Bresnahan M, Haugen M, Hornig M, Hirtz D, Lie K, Lipkin WI, Magnus P, Reichborn-Kjennerud T, Schjølberg S, Davey Smith G, Øyen A-S, Susser E, Stoltenberg C (2013) Association between maternal use of folic acid supplements and risk of autism in children. JAMA 309(6):570–577

Susser ES, Lin SP (1992) Schizophrenia after prenatal exposure to the Dutch Hunger Winter of 1944–1945. Arch Gen Psychiatry 49(12):983–988

Susser E, Shrout PE (2010) Two plus two equals three? Do we need to rethink lifetime prevalence? Psychol Med 40(6):895–897

Susser M, Susser E (1996) Choosing a future for epidemiology: I. Eras and paradigms. Am J Public Health 86(5):668–673

Susser E, Schwartz S, Morabia A, Bromet EJ (2006) Psychiatric epidemiology: searching for the causes of mental disorders. Oxford University Press, New York

Susser E, Baumgartner JN, Stein Z (2010) Commentary: Sir Arthur Mitchell – pioneer of psychiatric epidemiology and of community care. Int J Epidemiol 39(6):1417–1425

Susser E, Kirkbride JB, Heijmans BT, Kresovich JK, Lumey LH, Stein AD (2012) Maternal prenatal nutrition and health in grandchildren and subsequent generations. Annu Rev Anthropol 41:577–610

Terris M (1964) Goldberger on pellagra. Louisiana State University Press, Baton Rouge, Louisiana

Tohen M, Bromet E, Murphy JM, Tsuang MT (2000) Psychiatric epidemiology. Harv Rev Psychiatry 8(3):111–125

Torrey EF, Bowler AE, Rawlings R, Terrazas A (1993) Seasonality of schizophrenia and stillbirths. Schizophr Bull 19(3):557–562

Tsuang MT, Dempsey GM (1979) Long-term outcome of major psychoses. II. Schizoaffective disorder compared with schizophrenia, affective disorders, and a surgical control group. Arch Gen Psychiatry 36(12):1302–1304

Tsuang M, Tohen M (2002) Textbook in psychiatric epidemiology. Wiley-Liss, New York

Tuchman RF, Rapin I, Shinnar S (1991) Autistic and dysphasic children. II: Epilepsy. Pediatrics 88(6):1219–1225

USDHHS (1999) Mental health: a report of the Surgeon General. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services, National Institutes of Health, National Institute of Mental Health, Rockville

Ustun TB (1999) The global burden of mental disorders. Am J Public Health 89(9):1315–1318

Ustun TB, Ayuso-Mateos JL, Chatterji S, Mathers C, Murray CJ (2004) Global burden of depressive disorders in the year 2000. Br J Psychiatry 184:386–392

van Os J, Driessen G, Gunther N, Delespaul P (2000) Neighbourhood variation in incidence of schizophrenia: Evidence for person-environment interaction. Br J Psychiatry 176:243–248

van Os J, Hanssen M, Bijl RV, Vollebergh W (2001) Prevalence of psychotic disorder and community level of psychotic symptoms: an urban-rural comparison. Arch Gen Psychiatry 58(7):663–668

Veling W, Hoek HW, Selten JP, Susser E (2011) Age at migration and future risk of psychotic disorders among immigrants in the Netherlands: a 7-year incidence study. Am J Psychiatry 168(12):1278–1285

Wells JE, Bushnell JA, Hornblow AR, Joyce PR, Oakley-Browne MA (1989) Christchurch Psychiatric Epidemiology Study, part I: methodology and lifetime prevalence for specific psychiatric disorders. Aust N Z J Psychiatry 23(3):315–326

Wing L, Gould J, Gillberg C (2011) Autism spectrum disorders in the DSM-V: better or worse than the DSM-IV? [Reports - Evaluative]. Res Dev Disabil 32(2):768–773

Xu B, Ionita-Laza I, Roos JL, Boone B, Woodrick S, Sun Y, Levy S, Gogos JA, Karayiorgou M (2012) De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. Nat Genet 44:1365–1369

Yeargin-Allsopp M, Rice C, Karapurkar T, Doernberg N, Boyle C, Murphy C (2003) Prevalence of autism in a US metropolitan area. JAMA 289(1):49–55

Zimmermann MB (2012) The effects of iodine deficiency in pregnancy and infancy. Paediatr Perinat Epidemiol 26(Suppl. 1):108–117

Epidemiology of Diabetes

60

Matthias B. Schulze and Frank B. Hu

Contents

M.B. Schulze (✉)
Department of Molecular Epidemiology, German Institute of Human Nutrition
Potsdam-Rehbruecke, Nuthetal, Germany

F.B. Hu
Departments of Nutrition and Epidemiology, Harvard School of Public Health, Boston, MA, USA

Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard
Medical School, Boston, MA, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_66,

60.1 Introduction

Diabetes is a heterogeneous group of metabolic diseases which share a common characteristic: hyperglycemia. Hyperglycemia develops as a result of various defects in insulin secretion by pancreatic beta-cells or from changes in insulin action. Either defect jeopardizes the delicate interplay between insulin secretion and insulin action which is essential for the maintenance of normoglycemia. Normal pancreatic beta-cells can adapt to changes in insulin action. For example, under regular circumstances, a decrease in insulin sensitivity in target organs such as muscle is compensated by upregulation of insulin secretion and, alternatively, a decrease in secretory function of beta-cells can be counterbalanced by increases in insulin sensitivity. This hyperbolic relationship between insulin secretion and insulin action is deteriorated in patients with diabetes, resulting in chronically higher than normal levels of glycemia (Stumvoll et al. 2005). Type 2 diabetes, the most common form of diabetes, is characterized by a relative loss of beta-cell function which leads to insufficient compensation for increased insulin resistance. Type 1 diabetes is characterized by an absolute deficiency of insulin secretion resulting from autoimmune destruction of pancreatic beta-cells. Due to the different pathophysiology and causative factors involved, a detailed description of the epidemiology of type 1 diabetes is beyond this chapter. Interested readers should refer to previous reviews (Ekoé et al. 2008; Maahs et al. 2010).

Knowledge about the possible causes of type 2 diabetes has greatly advanced over the past few decades largely due to findings from experimental and observational prospective studies in diabetes epidemiology. Unhealthy diet and lifestyle are the most important modifiable factors for type 2 diabetes, and observational studies and randomized clinical trials have demonstrated that diabetes is largely preventable through lifestyle interventions. Accordingly, several countries have responded by implementing national diabetes programs (Colagiuri et al. 2010) and developing guidelines for diabetes prevention (Paulweber et al. 2010). However, despite our improved understanding of the determinants for type 2 diabetes, obesity and physical inactivity have continued to grow in most parts of the world. Particularly in developing countries that are experiencing an economic transition, the current diabetes trend is exceedingly alarming and deserves immediate attention.

The following section describes criteria for diagnosing diabetes and the rationale behind diagnostic cut-offs; the third section discusses current estimates of prevalence and incidence; the fourth section lays out common methodological approaches in diabetes epidemiology; the fifth section presents evidence for different lifestyle risk factors; biochemical and genetic biomarkers as risk factors for type 2 diabetes are discussed in Sects. 60.6 and 60.7, respectively; Sect. 60.8 summarizes evidence from observational and experimental studies on the preventability of type 2 diabetes; Sect. 60.9 discusses approaches for screening to detect undiagnosed cases and high-risk individuals; and the final section draws some conclusions from the chapter.

60.2 Diagnostic Criteria of Diabetes

Classic symptoms of diabetes are polyuria, dehydration, and increased thirst due to glycosuria. Glycosuria occurs if the glucose concentration in the blood is raised beyond the renal threshold, which may vary substantially between individuals (Butterfield et al. 1967). The diagnostic criteria applied today with regard to fasting plasma glucose concentrations are considerably lower than the renal threshold (about 162–180 mg/dL/[9–10 mmol/L]) for most patients (Table 60.1).

All diagnostic criteria for diabetes are dependent on a threshold value that is not clearly linked to symptoms but instead are based on a continuous distribution of blood glucose values. Justification of the diagnostic thresholds has been a long-standing matter of debate. The original approach defined abnormal glucose values as those greater than the mean + 2 SD in a given population. This approach has not been found to correlate well with symptoms (Siperstein 1975). An alternative statistical approach, to be used if the distributions of glucose values follow a bimodal distribution with clear separation of diabetes patients and diabetes-free individuals, is the use of the anti-mode (McCance et al. 1994). This approach, however, is dependent upon population characteristics and is, therefore, unlikely to yield generalizable thresholds (Barr et al. 2002). The clinical approach widely used today is based on the assumption that there exists a threshold above which diabetic complications occur at an increased rate. Chronic exposure to hyperglycemia as a result of diabetes is associated with long-term damage, dysfunction, and failure of different organs, particularly the eyes, kidneys, nerves, heart, and blood vessels. However, diagnosing diabetes based on the relation of glucose values to diabetes complications remains problematic. Although patients with diabetes are at a two- to

Table 60.1 Diagnostic criteria of diabetes and prediabetes from the American Diabetes Association and the world Health Organization

	American Diabetes Association (2011)		World Health Organization/ International Diabetes Federation (2006, 2011)	
	Diabetes	Prediabetes	Diabetes	Prediabetes
Fasting plasma glucose	≥126 mg/dL (7.0 mmol/L)	100–125 mg/dL (5.6–6.9 mmol/L)	≥126 mg/dL (7.0 mmol/L)	110–125 mg/dL (6.1–6.9 mmol/L)
2h plasma glucose after ingestion of a 75-g glucose load	≥200 mg/dL (11.1 mmol/L)	140–199 mg/dL (7.8–11.0 mmol/L)	≥200 mg/dL (11.1 mmol/L)	140–199 mg/dL (7.8–11.0 mmol/L)
Hemoglobin A1c	≥6.5%	5.7–6.4%	≥6.5%	–

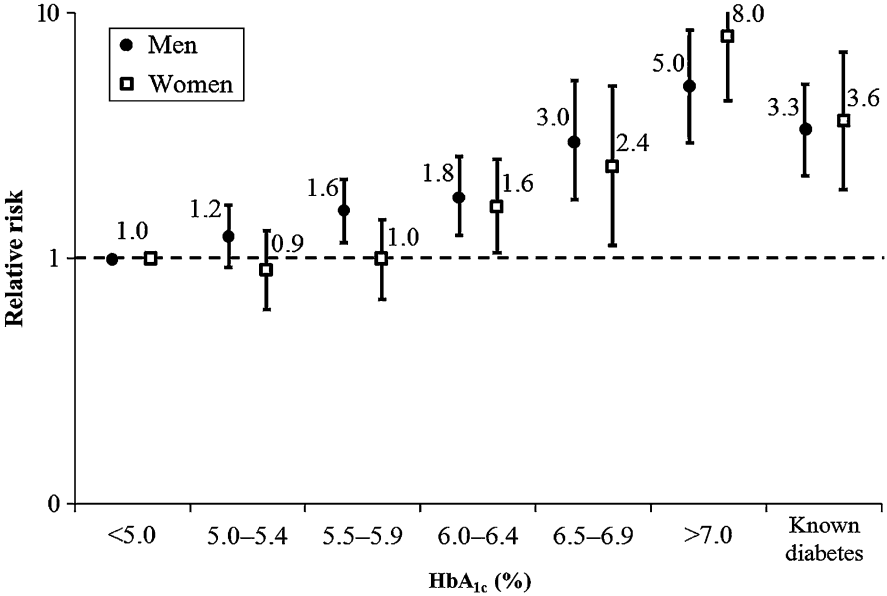


Fig. 60.1 Age-adjusted relative risk of cardiovascular disease events by categories of HbA_{1c} concentrations and known diabetes, EPIC-Norfolk Study (Khaw et al. 2004)

four-fold increased risk of coronary heart disease (Haffner and Cassells 2003; Hu 2002), there appears to be a relationship between glycemia and cardiovascular risk that extends well into the area that is clinically considered non-diabetic (Levitan et al. 2004; Sarwar et al. 2010). This association appears to be more evident when using hemoglobin A_{1c} (HbA_{1c}) compared to 2h plasma glucose (2hPG) during an oral glucose tolerance test or fasting plasma glucose (FPG) values (Sarwar et al. 2010). In the EPIC-Norfolk Study (Khaw et al. 2004), the largest prospective study which evaluated HbA_{1c}, there was a stepwise increase in risk for CVD with higher HbA_{1c} values over a follow-up period of 8 years. A 60–80% increase in risk was shown for HbA_{1c} values between 6.0% and 6.4% compared to participants with HbA_{1c} <5% (Fig. 60.1). Thus, the relationship between glycemia and cardiovascular risk does not provide evidence to substantiate a diagnostic threshold for diabetes.

In contrast to macrovascular complications, the associations between different measures of glycemia and the prevalence of microvascular complications are markedly non-linear (Barr et al. 2002). Still, defining optimal diagnostic thresholds from the shape of the associations has been challenging. For example, recent data from the DETECT-2 study (Fig. 60.2), a consortium of several studies with ~45,000 participants, suggest that the risk of diabetes-specific retinopathy increases around an HbA_{1c} of 6.5% and a 2hPG in the range of 10.1–11.2 mmol/L. These findings support current diagnostic thresholds (Colagiuri et al. 2011). However, the optimal

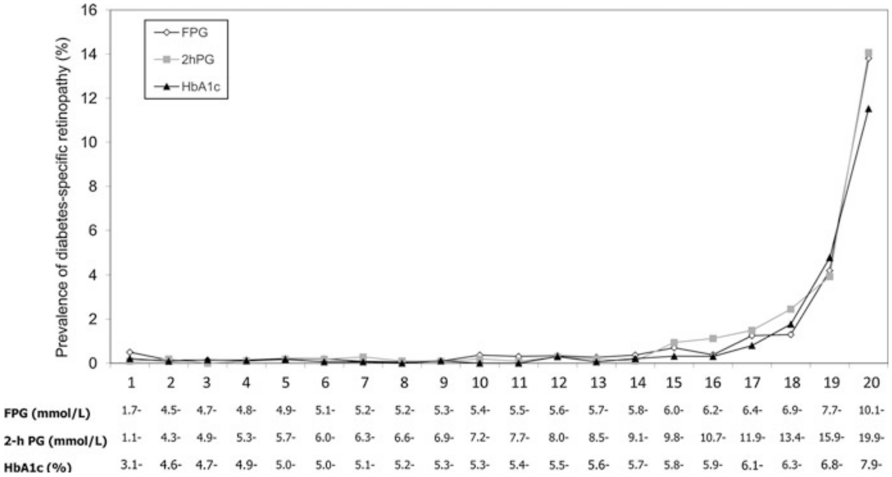


Fig. 60.2 Prevalence of diabetes-specific retinopathy (moderate or more severe retinopathy) by vigintiles of the distribution of FPG, 2hPG, and HbA$_{1c}$ (Colagiuri et al. 2011)

cut-off for FPG in the DETECT-2 study was found to be around 6.5 mmol/dL (117 mg/dL), which is considerably lower than the current diagnostic cut-off of 7.0 mmol/dL (126 mg/dL).

60.3 Prevalence and Incidence of Diabetes Across the World

According to estimations of the International Diabetes Federation, the global prevalence of diabetes in 2010 was 6.4% for adults between 20 and 79 years of age (International Diabetes Federation 2009). In absolute terms, approximately 285 million people in this age range are estimated to have diabetes. There are substantial regional differences in diabetes prevalence. The highest age-adjusted prevalence has been observed in the North American and Caribbean region (10.2%) and the Middle East and North African region (9.3%) (Table 60.2). Global projections predict that the total number of affected individuals will continue to increase over the coming years. The International Diabetes Federation estimates that by the year 2030, 438 million people will have diabetes, accounting for 7.7% of the total population in this age range. These projections probably underestimate the impact of the diabetes epidemic as they are only based on the expected population growth and age-specific demographic changes as determinants of disease prevalence, without taking changes in the prevalence of behavioral and lifestyle risk factors into account. Large differences in diabetes prevalence exist when people from rural and urban geographical areas (with the same or similar ethnicity) are compared. This indicates that the ongoing transition in behavioral, environmental, economic, and social risk factors (urbanization, unhealthy diets, physical inactivity, obesity) will most likely

Table 60.2 Worldwide prevalence of diabetes in 2010 and 2030 among adults (20–79 years) (International Diabetes Federation 2009)

	2010		2030	
Region	Population (millions)	Diabetes prevalence[a] (%)	Population (millions)	Diabetes prevalence[a] (%)
Africa	379	3.8	653	4.7
Europe	646	6.9	659	8.1
Middle East and North Africa	344	9.3	533	10.8
North America and Caribbean	320	10.2	390	12.1
South and Central America	287	6.6	382	7.8
Southeast Asia	838	7.6	1,200	9.1
Western Pacific	1,531	4.7	1,772	5.7
Total	4,345	6.4	5,589	7.7

[a] Prevalence adjusted to world population in 2010 and 2030

result in accelerated growth in diabetes prevalence in many parts of the world, particularly in developing countries.

Representative data on type 2 diabetes incidence rates are lacking from most parts of the world. Many prospective cohort studies collect diabetes incidence data; however, participants are usually selected based on specific geographical regions, age ranges, or individual characteristics. Thus, prospective cohort studies are frequently conducted in populations that are not nationally representative. Countries with nationally representative studies generally draw from cross-sectional survey data, medical records, registries, or health administration data to reflect the current incidence of diagnosed diabetes. For example, estimates of diabetes incidence in the US are based on the representative National Health Interview Survey. The survey data are continuously available; however, the survey does not distinguish whether increases in diabetes incidence rates are due to an actual increase in the number of cases or to improved case ascertainment or a combination of these factors. Based on this survey, the incidence of diagnosed diabetes has increased in the USA from 4.9 to 6.9 per 1,000 adults (18–79 years) between 1997 and 2003 (Geiss et al. 2006). These data, in combination with trend data on the ratio of diagnosed and undiagnosed diabetes (Gregg et al. 2004), provide strong evidence that diabetes incidence has substantially increased during recent years in the USA, likely due to increased obesity and changes in behavioral risk factor prevalence.

There are several methodological difficulties in the assessment of diabetes prevalence and incidence which hinder the comparability of country- or region-specific estimates (Table 60.3). For example, the data sources used by the International Diabetes Federation in determining the country-specific prevalence rates draw from considerably heterogeneous study populations and diabetes assessment techniques (International Diabetes Federation 2009). While representative studies are likely to produce reliable prevalence estimates, such studies are not available for many

Table 60.3 Methodological problems in estimation of diabetes prevalence and incidence (Schulze et al. 2010a)

Problem	Comment
Unknown diabetes	• Clinical diagnosis is made frequently by chance • Considerable proportion of cases are likely unknown • Proportion of unknown cases depends on the availability and acceptability of population-wide screening
Diagnostic parameters	• Guidelines offer a variety of alternative parameters with changes over time (e.g., recent inclusion of HbA_{1c}) • Different parameters do not identify similar case groups • Prevalence of known diabetes in a population depends on the parameters commonly used in clinical practice
Diagnostic thresholds	• Changes in diagnostic thresholds affect prevalence estimates (e.g., increasing prevalence after lowering the threshold for FPG from 140 to 126 mg/dL)
Reliability of measures	• Clinical diagnosis requires confirmatory measurement • Reproducibility of elevated glycemia levels: ~50% • Reproducibility is different for different parameters (e.g., higher for FPG than for 2hPG) and depending on time

countries. Frequently, data from selected study populations are not generalizable to the national level. For countries where representative studies have been conducted, a difference in time intervals or diagnostic parameters makes cross-country comparisons challenging.

Universal glucose screening has not been common practice in most countries, and insurance coverage of screening programs is frequently low. Thus, individuals meeting diagnostic criteria may remain undetected in the general population at any point in time. For example, in a representative study in a region of southern Germany, the prevalence of undetected diabetes has been estimated to be as high as the prevalence of diagnosed diabetes (Meisinger et al. 2010; Rathmann et al. 2003). In the USA, undiagnosed diabetes cases account for about 1/5 of all diabetes cases in NHANES 2003–2006 (overall prevalence among adults $\geq$20 years of age is 9.6%) (Cowie et al. 2010). To estimate the total population level of prevalence and incidence, representative studies would need to identify all individuals previously diagnosed with diabetes as well as those currently meeting the diagnostic criteria. Data from the US-based NHANES 2005–2006 indicate that there is limited overlap across the three diagnostic parameters FPG, 2hPG, and HbA_{1c} (Cowie et al. 2010). Of the three diagnostics, 2hPG results in the highest prevalence of undiagnosed diabetes (4.9%), followed by FPG (2.5%) and HbA_{1c} (1.6%).

The actual proportion of undiagnosed diabetes in a population is a function of the true total prevalence and the extent to which diabetes cases are identified within the health care system through screening programs. In most cases, survey estimates for the prevalence of undiagnosed diabetes rely on a single diagnostic measurement. Studies have suggested that about half of the cases detected through screening utilizing 2hPG as the diagnostic criterion will be confirmed in a second measurement (Brohall et al. 2006; Eschwege et al. 2001). The results from the 2hPG diagnostic are less reproducible than the results from FPG or HbA_{1c} screening. Thus, diabetes prevalence is overestimated when glucose parameters are measured only once. To increase accuracy, repeated measurements that confirm screen-detected cases are necessary. Logistically, this clearly presents a challenge.

60.4 Approaches in Analytical Diabetes Epidemiology

While descriptive diabetes epidemiology describes the distribution, prevalence, and incidence of diabetes in populations, analytical epidemiology investigates the determinants of diabetes development. Observational prospective cohort studies have been widely used in diabetes epidemiology to investigate the determinants of disease (for a detailed description of cohort studies see chapter ▸Cohort Studies of this handbook). In addition, several randomized controlled trials have been conducted to evaluate the efficacy of lifestyle interventions and drug treatments in the prevention of diabetes. Prospective cohort studies are less prone to reverse causation and information bias (e.g., recall bias) than case-control or cross-sectional studies. Therefore, prospective cohort studies are considered the strongest study design among observational studies. Nested case-control studies and case-cohort studies (see chapter ▸Modern Epidemiological Study Designs of this handbook) can benefit from the efficient use of existing biological samples collected through prospective cohort studies. Although randomized clinical trials can provide the strongest evidence for causal inference, such trials are often infeasible for individual dietary and lifestyle factors due to high cost and lack of long-term compliance.

Lifestyle factors play an important role in the etiology of type 2 diabetes (see Sect. 60.5). Thus, methodological developments in assessment tools are essential to diabetes epidemiology. Optimal methods for collecting data on diet and physical activity have been a longstanding source of debate (see chapters ▸Nutritional Epidemiology and ▸Physical Activity Epidemiology of this handbook). Semiquantitative food frequency questionnaires are the most commonly used method to assess diet in nutritional epidemiological studies, but the validity of this method varies across different populations. To account for the correlations between intakes of energy and nutrients, adjustment of dietary data requires sophisticated statistical modeling (see chapter ▸Nutritional Epidemiology of this handbook). In addition, anthropometric measures such as weight, height, and waist and hip circumferences have been commonly used to assess body fat and fat distribution in epidemiological studies. These measures tend to have a high degree of intercorrelations, and simultaneous modeling of these variables is subject to various interpretations (see Sect. 60.5.1). Complex modeling of risk factors typically employs logistic (see chapter ▸Regression Methods for Epidemiological Analysis of this handbook) or Cox (see chapter ▸Survival Analysis of this handbook) regression models.

60.5 Major Lifestyle Risk Factors

60.5.1 Overweight and Obesity

Excessive body fat is the single largest risk factor for type 2 diabetes. For a detailed review on potential mechanisms by which obesity leads to insulin resistance and type 2 diabetes, the reader should refer to Kahn et al. (2006). The diabetes risk associated with excessive body fat, measured by the body mass index (BMI, the ratio of body weight in kg to squared height in meter) or anthropometric indicators such as waist circumference or skinfold thickness, increases in a continuous fashion. Clinical risk categories for BMI (normal weight 18.5–24.9 kg/m^2, overweight 25–29 kg/m^2, and obesity $\geq$30 kg/m^2) are associated with a stepwise increase in diabetes risk. However, studies have clearly shown that diabetes risk increases already within the normal body weight range (Hu et al. 2001a). Existing evidence from randomized controlled trials has convincingly demonstrated the benefits of weight reduction on diabetes incidence, but these studies were limited to high-risk individuals who were overweight (Knowler et al. 2002; Ramachandran et al. 2006; Tuomilehto et al. 2001).

Whether anthropometric measures that reflect body fat distribution are superior to measures of total or percent body fat has been a matter of debate. A meta-analysis of prospective observational studies suggests that the relative risk associated with higher waist circumference is slightly stronger than that associated with higher BMI (Vazquez et al. 2007). These findings suggest that waist circumference is a valid alternative to BMI when assessing type 2 diabetes risk in a clinical setting or at the population level, although the combination of BMI and waist circumference can be more predictive of diabetes risk. Waist-height ratio has also been investigated in a number of studies, although the added value beyond waist circumference alone remains unclear (Browning et al. 2010; Schulze et al. 2006; Taylor et al. 2010). Determining whether abdominal adiposity predicts type 2 diabetes independently of general adiposity has become a research priority. Numerous studies have used multivariable modeling to examine the role of body fat indices, but the colinearity of anthropometric measures is analytically challenging, and the interpretation of mutually adjusted body fat measurements is not straightforward (Table 60.4).

While BMI is the most frequently used clinical and epidemiological measure of body fat and is generally thought to be uncorrelated with height, studies indicate that BMI is slightly correlated with height in many populations, particularly among women (Diverse Populations Collaborative Group 2005). Height and diabetes risk have an inverse relationship (Schulze et al. 2006; Weitzman et al. 2010). Similarly, waist and hip circumferences are correlated with height (Heymsfield et al. 2011). These correlations have implications not only for defining clinical cut-offs for waist and hip circumferences but also for estimating the strength of the association between body fat measures and diabetes risk. Waist-hip circumference measures reflect fat accumulation in the abdominal region more accurately than BMI, but these measures are still imprecise as they are not specific enough to assess the

Table 60.4 Conceptual meanings of statistical models using various anthropometric variables to predict type 2 diabetes risk (Hu 2008)

Models	Interpretations
1. $Y =$ height + weight	Coefficient for weight can be interpreted as the association between overall body fatness and disease risk; the interpretation of height is unclear as, to some degree, it becomes a surrogate for lean body mass
2. $Y =$ height + BMI	BMI and height are uncorrelated. BMI is a measure of overall adiposity, while height can be interpreted as a surrogate of childhood and adolescent nutritional status
3. $Y =$ height + weight adjusted for height	The correlation between height and weight adjusted for height (residuals) from a regression model is zero. Weight adjusted for height is a marker of overall adiposity, while height is a surrogate of childhood and adolescent nutritional status
4. $Y =$ BMI + WC (or WHR)	BMI and WC (or WHR) are highly correlated. While WC is a measure of central obesity, the interpretation of BMI (holding WC constant) is complicated, as it largely reflects the effects of muscularity rather than body fatness, especially in the elderly
5. $Y =$ BMI + WC adjusted for BMI	BMI and WC adjusted for BMI (residuals) in a regression model is zero. WC residuals represent the effects of central obesity adjusted for overall adiposity, while BMI represents the effect of overall adiposity
6. $Y =$ WC + hip circumference	Waist and hip circumferences are moderately correlated. While WC is a measure of central obesity or abdominal fat, hip circumference (holding WC constant) largely represents the effects of gluteal muscularity and bone structure
7. $Y =$ baseline weight + current weight	After adjusting for baseline weight, current weight largely reflects the effects of change in body weight on disease risk
8. $Y =$ change in weight + change in WC	change in WC represents the effects of changes in body fat distribution on disease risk, while change in weight (holding change in WC constant) largely reflects changes in lean body mass (e.g., in the elderly, weight loss is largely due to muscle loss)

Y disease outcome (e.g., type 2 diabetes), *BMI* body mass index, *WC* waist circumference, *WHR* waist-to-hip circumference ratio

amount of visceral and subcutaneous fat. Studies have shown that visceral adipose tissue is metabolically more active than subcutaneous adipose tissue (Jain et al. 2009).

60.5.2 Physical Activity

The notion that physical activity is a central element in diabetes prevention is supported by evidence from multiple major lifestyle intervention trials (Knowler et al. 2002; Ramachandran et al. 2006; Tuomilehto et al. 2001), although it is difficult to disentangle the independent roles of physical activity and dietary interventions in these trials. In the Da Qing Study, a group-randomized trial of high-risk individuals with impaired glucose tolerance conducted in China, physical activity intervention alone had a significant, beneficial impact on diabetes incidence in comparison with the standard intervention (Pan et al. 1997).

A large body of observational studies further supports the beneficial role of physical activity in diabetes prevention (Wareham 2007). Physical activity is a cornerstone of weight maintenance, and it is associated with increased insulin sensitivity (Maarbjerg et al. 2011). Several prospective studies have demonstrated a reduction in diabetes risk with higher levels of physical activity (reviewed in Gill and Cooper (2008) and Wareham (2007)). In most studies, a significant inverse association between physical activity and diabetes remained after adjusting for BMI, suggesting that the benefits of physical activity on diabetes are not entirely mediated through body weight.

Measuring physical activity by questionnaires is prone to measurement errors in epidemiological studies (see chapter ▸Physical Activity Epidemiology of this handbook), likely resulting in an underestimation of the true effect size. Thus, the amount of physical activity required to prevent diabetes remains unresolved. New technologies of activity assessment, e.g. heart rate monitoring and accelerometers, provide more accurate estimates of physical activity levels, although the applications of these technologies in large populations are expensive and logistically difficult. Nonetheless, the combination of self-reported and objectively measured physical activity data will provide further insights on the beneficial role of different amounts and intensities of physical activity in the development of diabetes.

The type of activity most strongly related to a reduction in diabetes risk remains unclear. Moderate to vigorous activity, including brisk walking, has consistently been related to lower diabetes incidence (Jeon et al. 2007). However, no study has examined the role of resistance training such as weight lifting versus aerobic exercise in diabetes risk. Sedentary behaviors, such as prolonged television watching, are associated with an increased risk of diabetes. This relationship is not explained by unhealthy eating patterns associated with television watching (Grøntved and Hu 2011). Appropriate control for different physical activities in analyses of a specific activity as exposure can be performed using isotemporal substitution models reflecting the displacement of time spent on different activities (Mekary et al. 2009). However, the multidimensional nature of physical activity (or inactivity) makes it hard to conclude whether the benefits are due to increased total energy expenditure (i.e., that any form of activity is advisable) or the fitness-producing effect of activities (i.e., that prevention efforts need to focus primarily on fitness-promoting recreational activities). Although several studies have suggested that higher fitness is associated with lower diabetes risk independently of body fatness (Carnethon et al. 2009; Lee et al. 2009a), the benefits of physical activity are likely due to the combination of both increased energy expenditure and improved physical fitness.

60.5.3 Smoking

Active cigarette smoking has consistently been associated with increased diabetes incidence (for a systematic review the reader can refer to Willi et al. (2007)).

Although the relative risk for active smokers compared to never smokers is moderate (∼1.5), smoking accounts for a considerable proportion of diabetes cases due to its high prevalence in many populations. In the USA, where smoking prevalence is declining, smoking has been estimated to account for about 12% of all diabetes cases (Ding and Hu 2007).

Smoking cessation is associated with a modest increase in weight. There is a dose-response relationship among former smokers with weight gain being proportional to the number of cigarettes formerly smoked daily (Filozof et al. 2004). However, this weight gain typically occurs in the shortterm. While diabetes risk is particularly high among individuals who recently quit smoking (Yeh et al. 2010), the beneficial effects of smoking cessation outweigh the adverse effects of associated weight gain in the longterm, leading to a reduction of diabetes risk (Wannamethee et al. 2001; Will et al. 2001).

60.5.4 Alcohol Consumption

Moderate alcohol consumption (1–3 drinks/day) has consistently been associated with lower diabetes incidence. Most studies have observed a u-shaped association, where an increased risk of adverse health outcomes is observed for abstainers and for heavy alcohol consumers. The large number of prospective observational studies that addressed alcohol consumption and diabetes risk has been summarized by several meta-analyses (Baliunas et al. 2009; Carlsson et al. 2005; Koppes et al. 2005). In most studies, the increased risk among alcohol abstainers compared to moderate drinkers should be interpreted cautiously because of heterogeneity in the abstainer population (e.g., life-long abstainers, sickquitters, underreporters). Nonetheless, the benefits of moderate alcohol consumption on diabetes still persisted even when light drinkers (e.g., half a drink or less per day) were used as the reference group.

For moderate alcohol consumers, the reduced risk of type 2 diabetes could be due to increased insulin sensitivity (Davies et al. 2002; Joosten et al. 2008) and adiponectin concentrations (Beulens et al. 2008a). However, results from randomized controlled trials are not entirely consistent, particularly for male participants (Beulens et al. 2006, 2007, 2008b; Sierksma et al. 2004; Zilkens et al. 2003). Although some studies observed different associations depending on the type of alcoholic beverage consumed, randomized trials suggest that the underlying biological mechanism is most likely to be explained by the consumption of alcohol, regardless of the type (Beulens et al. 2006, 2007, 2008b; Davies et al. 2002).

60.5.5 Dietary Factors

It has been hypothesized that higher total fat intake contributes to diabetes, directly by inducing insulin resistance and indirectly by promoting weight gain. Results from metabolic studies in humans, however, are inconsistent and generally do not support the idea that high-fat diets have a detrimental effect on insulin sensitivity (Lichtenstein and Schwab 2000; Riserus et al. 2009). Additionally, changing the dietary fat and carbohydrate composition in intervention studies generally had no effect on subsequent weight change over a long-term period (Howard et al. 2006). In most observational prospective studies, total fat or carbohydrate intake was not associated with diabetes risk (for an overview of the literature, the reader should refer to reviews on this topic (Hu et al. 2001b; Melanson et al. 2009; Schulze and Hu 2005)). This point is further supported by the Women's Health Initiative, a large randomized trial in which women who consumed a low-fat diet had similar diabetes incidence rates compared with women who consumed a standard USA diet (Tinker et al. 2008). The specific type of fat and carbohydrate may be more important than the total intake. Prospective studies suggest that diets that favor plant fats over animal fats (Hu et al. 2001b; Melanson et al. 2009) and that are rich in fiber, particularly from cereals (Schulze et al. 2007), are advantageous.

Carbohydrate quality can be determined by evaluating the physiologic response to carbohydrate-rich foods. The glycemic index reflects the quality of carbohydrates by ranking the ability of specific foods to raise postprandial blood glucose levels (Jenkins et al. 1981), whereas the glycemic load, a crossproduct of the glycemic index of a specific food and the amount of carbohydrates, reflects both quality and quantity of the carbohydrates. The relationship between glycemic index or load and risk of diabetes has been evaluated by a number of prospective studies, which showed that diets with low average glycemic index and glycemic load might be associated with lower risk for diabetes compared with high glycemic index/load diets (Dong et al. 2011a; Liu and Chou 2010). These associations appear to be independent of the amount of dietary fiber.

Though data are limited with regard to dietary protein, higher intake of animal protein has been observed to be associated with higher diabetes risk (Schulze et al. 2008; Sluijs et al. 2010; Song et al. 2004). In addition, cohort studies suggest that higher magnesium intake (Dong et al. 2011b; Larsson and Wolk 2007; Schulze et al. 2007) decreases diabetes risk, whereas higher iron intake – particularly from animal sources (iron bound to heme in oxygen-binding proteins myoglobin and hemoglobin) – increases diabetes risk (Rajpathak et al. 2009a). The relationship between antioxidative nutrients and diabetes risk has been evaluated in post hoc analyses of several randomized clinical trials. Generally, no benefits from vitamin C, vitamin E, or beta-carotene supplementation were found (Song et al. 2009), but supplementation with selenium was associated with increased risk of diabetes in one trial (Stranges et al. 2007).

A growing number of studies have evaluated the role of specific foods with regard to diabetes risk. Coffee consumption has been associated with lower diabetes risk in a large number of studies (Huxley et al. 2009; van Dam and Hu 2005). Residual confounding is unlikely to explain these results because regular coffee consumption is generally associated with unfavorable lifestyle habits in most populations. Whole grains have consistently been associated with lower diabetes risk in prospective studies (de Munter et al. 2007; Priebe et al. 2008). Similarly, dairy consumption is associated with lower diabetes risk (Tong et al. 2011), although this benefit may be restricted to low-fat dairy products. Consumption of nuts has also been associated with lower risk of diabetes, although to date, very few prospective studies have directly evaluated this hypothesis (Kendall et al. 2010). In contrast, frequent consumption of red and processed meats has consistently been related to higher diabetes risk in prospective cohort studies (Aune et al. 2009; Micha et al. 2010; Pan et al. 2011). Prospective studies also suggest that consuming sugar-sweetened beverages increases the risk of developing type 2 diabetes (Malik et al. 2010). A meta-analysis indicates that higher consumption of fruits and vegetables is not significantly associated with diabetes risk (Carter et al. 2010), although green leafy vegetables were found to be protective. It should be noted that observational studies may not be able to capture the true effect of diet on disease risk due to measurement errors inherent in the use of questionnaires, which could potentially lead to an underestimation of the effect (Harding et al. 2008).

To capture an individual's exposure to overall diet, several methods have been developed to derive dietary patterns. For details on food pattern evaluation methods and studies that evaluate the relationship between dietary patterns and diabetes risk, the reader should refer to published reviews (Esposito et al. 2010; Kastorini and Panagiotakos 2009; Michels and Schulze 2005; Schulze and Hu 2002; Schulze and Hoffmann 2006). Studies that evaluate major eating patterns through the use of exploratory patterning methods (e.g., factor and cluster analysis) suggest that "Western" diets rich in red and processed meats, sugary drinks, and refined grains are related to higher diabetes risk (Fung et al. 2004; van Dam et al. 2002). Several authors have used the reduced rank regression technique, which allows the use of intermediate risk markers in pattern recognition to identify diabetes-related patterns (Heidemann et al. 2005; Imamura et al. 2009; Liese et al. 2009; McNaughton et al. 2008; Schulze et al. 2005). Most studies support the notion that dietary patterns that favor fruits, vegetables, whole grains, and vegetable fats at the expenses of red meats, refined grains, and sugared soft drinks reduce the risk of type 2 diabetes. While this evidence is observational, additional support for the beneficial effects of similar diet patterns (e.g., the Mediterranean diet) comes from intervention studies (Salas-Salvado et al. 2011).

Measuring dietary intake in observational studies has been a major focus of research in nutritional epidemiology (see chapter ▶Nutritional Epidemiology of this handbook). Semiquantitative food frequency questionnaires are the most commonly used method to assess diet in large-scale studies to date. Food frequency questionnaires can be self-administered by participants, are relatively easy to complete, can be processed by computer, and are inexpensive – features that make them particularly feasible for use in large epidemiological studies. However, the validity of this method may vary with different populations and cultures. In addition, because food frequency questionnaires lack the detail and specificity of diet records or recalls, they may not provide accurate estimates of absolute intake of some nutrients. As a consequence, food frequency questionnaires provide very useful information on relative ranking of usual nutrient and food intakes rather than precise amounts of intakes. While more quantitative methods (food records, 24h-recalls) are available, they are short-term assessment instruments unlikely to reflect long-term usual intake and put considerable burden on study participants and investigators. New methodological developments focus on the combination of different assessment methods. For example, data from food frequency questionnaires can be used in combination with few repeated 24-h recalls to increase the validity of collected dietary data (Souverein et al. 2011).

In epidemiological studies on diet and diabetes, the regression calibration approach can be used to correct measures of association for random and systematic errors inevitable in all dietary assessments (Qiu and Rosner 2010). However, this approach requires a validation study in a subsample of the study. Also, random measurement errors can also be reduced by repeated dietary assessment over the course of a study (Hu et al. 1999). For example, correction for measurement error by the regression calibration approach and by using repeated dietary assessments in comparison to baseline diet only yielded stronger associations between red and processed meat intake and type 2 diabetes risk (Pan et al. 2011).

Due to the multidimensional nature of dietary intake, observational studies on dietary risk factors and diabetes require careful statistical model building. Given that several dietary exposures have been observed to be related to diabetes risk, confounder adjustment is a central requirement. Besides control of confounding, adjustment for total energy intake also mimics differences in dietary composition under isocaloric situations. In the case that the effect of specific energy-providing macronutrients on disease outcome is modeled, effect estimates then reflect macronutrient substitutions. For example, adjustment of carbohydrate density (percentage of total energy) for total energy intake and specific other macronutrient densities allows to evaluate associations with diabetes risk for an isocaloric substitution of carbohydrates for fat or for protein (Schulze et al. 2008). Multivariable modeling can also be used to investigate the effect of a substitution of a serving of one food for another (Halton et al. 2006; Pan et al. 2011).

60.5.6 Emotional Stress and Sleeping Problems

Findings from prospective cohort studies suggest that different forms of emotional stress increase the risk of type 2 diabetes. Here, emotional stress refers to consequences of the failure to respond appropriately to emotional threats, with signs of stress defined at a cognitive, emotional, physical, or behavioral level (Pouwer et al. 2010). Depression has consistently been associated with higher diabetes incidence (Knol et al. 2006; Mezuk et al. 2008). Also, general emotional stress has been found to increase diabetes risk, although the data are not conclusive (Pouwer et al. 2010). There is some evidence that psychosocial stress at work is also associated with increased diabetes risk (Agardh et al. 2003; Heraclides et al. 2009). Emotional stress can affect sleep duration and sleep quality. Conversely, sleeping problems may not only be a consequence of emotional stress but are often experienced as a significant source of stress (Pouwer et al. 2010).

Prospective studies have consistently observed higher diabetes risk with short sleep duration ($\leq$5–6 h/night) and long sleep duration ($>$8–9 h/night). There is also strong evidence that reduced sleep quality, e.g., difficulties in initiating or maintaining sleep, increases diabetes risk (Cappuccio et al. 2010). Potential mechanisms relating short sleep duration to increased risk include changes in circulating levels of leptin and ghrelin as well as cortisol metabolism and inflammation. While BMI is an important potential confounder because it can contribute to snoring problems and sleep apnea (and thus to sleeping problems), short sleep duration has generally been found to increase diabetes risk independently of body fatness. There is less clear indication of possible mechanisms mediating the risk-increasing effect of long sleep duration, although confounding by depressive symptoms, low socioeconomic status, unhealthy lifestyle patterns, and prevalent health conditions may be potential explanations (Pouwer et al. 2010).

60.6 Biochemical Predictors

In light of the increasing prevalence of diabetes worldwide, interest in identifying "novel" predictors is mounting. A growing number of studies have evaluated biochemical markers for associations with diabetes risk in recent years. Such research has mainly been targeted at improving our understanding of the pathogenesis of diabetes. Biochemical markers have also been increasingly used to evaluate potential mechanisms by which conventional risk factors might be related to diabetes risk. These investigations focus on etiology and can be separated from the debate on whether biomarkers have value as a screening tool for predicting future diabetes cases. The latter point is discussed later in this chapter.

The evaluation of biochemical markers in epidemiological studies involves a number of challenges which are often not sufficiently addressed (see chapter ►Molecular Epidemiology of this handbook). These complex problems include: residual confounding, inappropriate adjustment for intermediate variables, and analytical measurement error and biological variation. For a more detailed discussion, the reader might refer to Sattar et al. (2008). While biochemical markers of diabetes risk may generally be measureable in different body tissues, the overwhelming majority of studies have focused on peptides, proteins, and metabolites measureable in peripheral blood. Parameters of glucose metabolism have been of particular interest because these are either directly relevant as diagnostic parameters (FPG, 2hPG, HbA_{1c}) or they reflect the primary underlying mechanisms of insulin resistance (reflected by different indices, e.g., HOMA insulin resistance index) and impaired β-cell function (reflected by fasting insulin, proinsulin, or HOMA β-cell function index). Additional biomarkers that have been found to predict diabetes incidence include blood lipids (i.e., triglycerides and HDL-cholesterol (Fagot-Campagna et al. 1997; Montonen et al. 2011; Schmidt et al. 2005; Stern et al. 2002; Wilson et al. 2007)), liver enzymes (alanine aminotransferase and gamma-glutamyltransferase (Fraser et al. 2009)) and other hepatic-derived predictors (sex hormone-binding globulin (Ding et al. 2006)), the adipokine adiponectin (Li et al. 2009), and markers of subclinical inflammation, in particular CRP (Lee et al. 2009b). There is also emerging evidence that novel markers related to inflammation and endothelial dysfunction (e.g., IL-6, cellular adhesion molecules, white cell count (Goldberg 2009)), body iron stores (Forouhi et al. 2007; Jehn et al. 2007; Rajpathak et al. 2009b), hepatokine fetuin-A (Ix et al. 2008; Stefan et al. 2008), or PAI-1 (Festa et al. 2002; Kanaya et al. 2006) are associated with diabetes risk. For a detailed review of the literature, the reader may refer to published reviews (Herder et al. 2011; Sattar et al. 2008).

Current approaches focus on complementing candidate-based biomarker studies with hypothesis-free approaches. In particular, metabolomics has gained more interest because metabolites are considered to be the most proximal biomarkers of pathophysiological processes. The number of different metabolites (including lipids, sugars, nucleotides, amino acids, organic acids), their substantial chemical diversity in terms of polarity and water/lipid solubility, and their wide range of concentration levels across different human tissues make the application of metabolomics in large-scale epidemiological studies particularly challenging (see chapter ▸Molecular Epidemiology of this handbook). A recent prospective study using a targeted metabolomics approach indicated that branched-chain and aromatic amino acids in plasma could be used as predictors of risk of type 2 diabetes (Wang et al. 2011).

60.7 Genetic Predictors

In addition to biochemical markers, there has been a growing interest in genetic markers as predictors of diabetes. Several loci were identified in candidate association studies (PPARG, KCNJ11, TCF7L2, WFS1, and HNF1B), but the majority of the known susceptibility loci for type 2 diabetes have been identified by genome-wide association studies (GWAS) based on case-control studies (e.g., FTO, SLC30A8, HHEX-IDE-KIF11, CDKAL1, IGF2BP2, CDKN2A-CDKN2B, TSPAN8, ADAMTS9, NOTCH2, CDC123-CAMK1D, THADA, JAZF, KCNQ1, IRS1, DUSP9, ZFAND6, PRC1, CENTD2, TP53INP1, KLF14, ZBED3, BCL11A, HNF1A, CHCHD9, HMGA2, UBE2E2, C2CD4A-C2CD4B, and RBMS1-ITGB6) or by evaluating associations with diabetes-related quantitative traits, e.g., FPG, 2hPG, HbA_{1c}, and indices of insulin resistance and β-cell function (MTNR1B, DGKB-TMEM195, GCKR, GCK, PROX1, and ADCY5) (for a detailed description of genome-wide association studies see chapter ▸Statistical Methods in Genetic Epidemiology of this handbook). GWAS have resulted in a major paradigm shift in epidemiological research from hypothesis-driven investigations toward exploratory analyses and decision-making based on the presence or absence of statistical significance. The pooling of study populations has led to increased statistical power and the discovery of many new loci. These loci are typically associated with a small to moderate effect on diabetes risk, with most variants carrying an odds ratio of

Table 60.5 Genetic variants evaluated to predict type 2 diabetes in prospective cohort studies considering multiple loci simultaneously

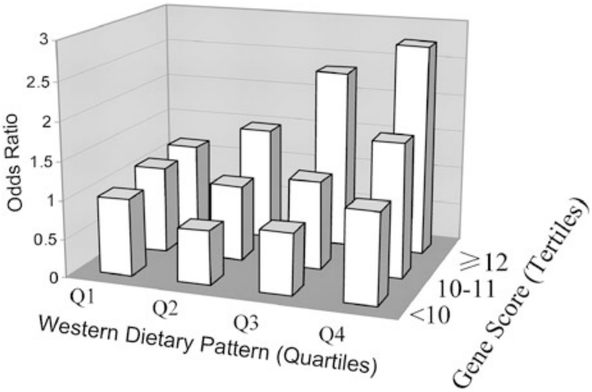
Study	Number of genetic variants	Comparison	Relative risk (95% CI)
Rotterdam Study (van Hoek et al. 2008)	18	Per risk allele	1.04 (1.02–1.07)
Framingham offspring Study (Meigs et al. 2008) (de Miguel-Yanes et al. 2011)	18 40	Per risk allele	1.12 (1.07–1.17) <50 years: 1.24 (1.13–1.36) ≥50 years: 1.11 (1.03–1.19)
Malmö preventive project (Lyssenko et al. 2008)	11	Per risk allele	1.12 (1.08–1.15)
Botnia Study (Lyssenko et al. 2008)	11	Per risk allele	0.94 (0.84–1.04)
Health professionals follow-up Study (Cornelis et al. 2009a)	10	Per risk allele	1.19 (1.14–1.24)
Nurses' health Study (Cornelis et al. 2009a)	10	Per risk allele	1.16 (1.12–1.20)
EPIC-Potsdam Study (Schulze et al. 2009)	20	≥22 vs. <17 risk alleles	1.47 (1.12–1.93)
Whitehall II Study (Talmud et al. 2010)	20	Per risk allele	1.7 (0.9–2.5)

1.1–1.2 per risk allele. In addition, these risk alleles are usually common, often exceeding 25% prevalence in most populations (Herder et al. 2011). Interestingly, the majority of risk loci reflect variation in genes involved in β-cell function rather than insulin resistance (Florez 2008).

Several recent prospective studies have used genetic risk scores or risk allele scores to capture genetic risk at multiple loci simultaneously (Table 60.5) (Cornelis et al. 2009a; de Miguel-Yanes et al. 2011; Lyssenko et al. 2008; Meigs et al. 2008; Schulze et al. 2009; Talmud et al. 2010; van Hoek et al. 2008). Such scores can generally be calculated by either assuming equal contribution of each risk allele or weighting the risk alleles. The latter technique has been applied by assigning study-specific or literature-based weights. Given that most genetic variants are associated with a modest increase in risk of diabetes and that study-specific associations are frequently inconsistent with associations reported from large pooling projects for single loci, the use of study-specific weights is likely to introduce overoptimism in risk prediction when evaluating genetic risk scores (de Miguel-Yanes et al. 2011). Also, because there is very little difference in effect sizes between most variants, weighted genetic risk scores may not be substantially different in their associated risk compared to that for unweighted scores (Cornelis et al. 2009a; Talmud et al. 2010).

The risk variants identified to date only account for ~10% of observed familial clustering of type 2 diabetes (Voight et al. 2010). Thus far, the identified variants

Fig. 60.3 Diabetes risk according to joint classification of a “Western” dietary pattern and genetic risk (Qi et al. 2009)



are common alleles (frequency >5%), and it remains unclear whether less common variants would have a stronger genetic effect on diabetes. Potentially, the combined effects of common and less common variants may explain a greater degree of heritability of diabetes, although rare variants with large effects are yet to be identified.

Current research is also focused on investigating the interplay between environmental risk factors and genetic susceptibility. For an overview of the literature, the reader should refer to reviews on this topic (Franks et al. 2007; Qi et al. 2008; Qi and Liang 2010). There is some evidence that the effectiveness of lifestyle interventions is dependent upon genetic variants (Florez et al. 2007). For example, prospective studies have observed that the beneficial effects of fiber-rich diets or diets with high carbohydrate quality (low glycemic index) may depend upon genetic variation in TCF7L2 (Cornelis et al. 2009b; Fisher et al. 2009). Additionally, the detrimental effects of adhering to a “Western” dietary pattern were stronger among individuals carrying a relatively large number of diabetes risk alleles compared to those with relatively few (Fig. 60.3) (Qi et al. 2009). Interactions between physical activity and genetic variants have also been found (Brito et al. 2009). Still, intensive lifestyle modification reduced diabetes risk irrespective of genetic susceptibility to diabetes in the Diabetes Prevention Program (Hivert et al. 2011). Although genetic predictors of diabetes have been identified by large consortia of cross-sectional and case-control studies in recent years, these studies are generally not suitable for evaluating gene-environment interactions because most of the studies did not collect exposure information on environmental factors, especially diet and lifestyle. Also, diet and lifestyle information assessed in case-control or cross-sectional studies is prone to recall bias and reverse causation. Analyses on gene-environment interactions require large prospective population-based studies with sufficient statistical power, detailed information on diet and lifestyle, and replication of the findings in various populations.

60.8 Prevention of Diabetes

60.8.1 Healthy Lifestyles and Preventability of Type 2 Diabetes

Epidemiological evidence strongly indicates that diabetes is associated with Western dietary and lifestyle habits. People who migrate to Western countries generally have more sedentary lifestyles and consume “Western” diets and, therefore, have a greater risk of developing type 2 diabetes compared with their counterparts who remain in their native countries (Manson and Spelsberg 1994). Populations undergoing westernization in the absence of migration have also experienced dramatic rises in obesity and type 2 diabetes (Gohdes et al. 1993; Collins et al. 1994; Hodge et al. 1994).

Although a large body of evidence from epidemiological studies has implicated individual dietary and lifestyle factors in the development of type 2 diabetes, only a few studies have examined multiple risk factors simultaneously (Ford et al. 2009; Gopinath et al. 2010; Hu et al. 2001a; Mozaffarian et al. 2009; Reis et al. 2011). All of these studies considered physical activity, diet, smoking, and overweight or obesity as modifiable risk factors (Table 60.6). Alcohol consumption was considered by three of the five studies as an additional risk factor (Hu et al. 2001a; Mozaffarian et al. 2009; Reis et al. 2011). Overall, participants who adhered to all of the low-risk behaviors had a dramatically lower risk of developing type 2 diabetes. The studies also observed that only a small fraction of participants fulfilled all low-risk behavior criteria and that the population attributable risk of not adhering to healthy lifestyles is very high.

In these studies, the definition of a healthy diet based on population percentiles is somewhat arbitrary, thus making the translation of findings into public health practice challenging. The categorization of continuous risk factors is also subjective as different cut-offs for BMI and waist circumference were used in different studies. Despite these limitations, the data provide strong epidemiological evidence that the majority of type 2 diabetes cases could be prevented by the adoption of a healthier lifestyle.

Several randomized trials have also demonstrated the preventability of diabetes through lifestyle modification (Table 60.7). In a group-randomized trial in China, intervention with diet alone, exercise alone, and diet-plus-exercise was associated with a 31%, 46%, and 42% reduction in the risk of developing diabetes compared to the control group (Pan et al. 1997). In the Finnish Diabetes Prevention Study, reduction in weight through dietary modification and increasing physical activity resulted in an overall diabetes risk reduction of 58% over 3 years (Tuomilehto et al. 2001). Similarly, in the US-based Diabetes Prevention Program, a lifestyle-modification program with the minimum goal of a 7% weight reduction and a minimum of 150 min of physical activity per week reduced diabetes incidence by 58% over 3 years (Knowler et al. 2002). In all of these studies, the effect was maintained several years beyond the active intervention period (Li et al. 2008; Lindstrom et al. 2006; Knowler et al. 2009). Similar effects of lifestyle modification on diabetes incidence have been observed in the Indian Diabetes Prevention

Table 60.6 Classification of low-risk, relative risk, and population attributable risk of type 2 diabetes mellitus in prospective studies for groups defined by combinations of modifiable risk factors

Lifestyle risk factors	Nurses' Health Study (Hu et al. 2001a)	Cardiovascular Health Study (Mozaffarian et al. 2009)	EPIC-Potsdam Study (Ford et al. 2009)	Blue Mountains Eye Study (Gopinath et al. 2010)	National Institutes of Health (NIH)–AARP Diet and Health Study (Reis et al. 2011)
Physical activity	Moderate-to vigorous exercise ≥30 min/day	≥median	≥3.5 h/week	≥3 times/week	20 min ≥ 3 times/week
Diet	Upper 2 quintiles of diet score	Upper 2 quintiles of diet score	>median of diet score	>median of diet score	Upper 2 quintiles of diet score
Smoking	No current	Never	Never	Never	Never or quit smoking >10 years ago
Overweight/obesity	BMI < 25 kg/m^2	BMI < 25 kg/m^{2}or waist circumference <88/92 cm	BMI < 30 kg/m^2	BMI < 30 kg/m^2	BMI 18.5–25 kg/m^2
Alcohol consumption	≥half a drink/day	Yes			Moderate
Percentage of population adhering to all factors	3.4%	3.4%	9.1%	11.4%	Men: 4.0% Women: 2.3%
Relative risk (95% CI)	5 factors vs. rest: 0.09 (0.05–0.17)	5 factors vs. rest: 0.11 (0.01–0.76)	4 factors vs. 0: 0.07 (0.05–0.12)	4 factors vs. 0: 0.17 (0.07–0.42)	5 factors vs. rest: Men: 0.28 (0.23–0.34) Women: 0.16 (0.10–0.24)
Population attributable risk % (95% CI)	91 (83–95)	89 (23–99)	Not reported	Not reported	Not reported

Table 60.7 Major lifestyle modification trials for prevention of type 2 diabetes

	Da Qing Study (Pan et al. 1997; Li et al. 2008)	Finnish Diabetes Prevention Study (Tuomilehto et al. 2001; Lindstrom et al. 2006)	Diabetes Prevention Program (Knowler et al. 2002, 2009)	Indian Diabetes Prevention Program (Ramachandran et al. 2006)	Japanese Trial (Kosaka et al. 2005)
N, sex	577 men and women with impaired glucose tolerance	522 men and women with impaired glucose tolerance	3,234 men and women with impaired glucose tolerance	269 men and women with impaired glucose tolerance	458 men with impaired glucose tolerance
Age	>25 years	40–65 years	>25 years	33–55 years	>30 years
Country	China	Finland	USA	India	Japan
Weight reduction goal	BMI = 23 kg/m^2 for overweight subjects in diet group	≥5%	7%	Weight maintenance	Reduction in BMI to <22 kg/m^2
Intervention arms	Diet alone Exercise alone Diet + exercise Control	Lifestyle intervention Control	Lifestyle intervention Control	Lifestyle intervention Control	Lifestyle intervention Control
Dietary intervention	25–30% energy from fat; 55–65% energy from carbohydrate; specific advice on cereal, vegetable, meat, milk, and oil intake in subjects with BMI ≥25 kg/m^2	<30% energy from fat; <10% saturated fat; fiber ≥15 g/1,000 kcal; frequent intake of whole grain products, vegetables, fruits, low-fat milk and meat products, soft margarines, and vegetable oils	25% fat, low-calorie diet; healthy eating based on US Department of Agriculture Food Guide Pyramid	Avoidance of simple sugars and refined carbohydrates; total fat intake ≤20 g/day; restriction of saturated fat; fiber-rich food	Smaller meals; avoidance of fat-rich foods
Lifestyle intervention	1 unit physical activity daily (5–30 min depending on intensity); 2 units/day if possible for those <50 years of age with no evidence of cardiovascular disease	Moderate exercise ≥30 min daily	150 min/week physical activity	Brisk walking for ≥30 min each day	Walking 30–40 min/day or cycling 30 min at weekends
Follow-up (years)	6	3.2	2.8	2.5	4
Diabetes risk reduction with intervention compared with controls	Diet only: 31% Exercise only: 46% Diet + Exercise: 42%	63%	58%	28%	67%

Program (Ramachandran et al. 2006) and in a Japanese trial (Kosaka et al. 2005). These trials share the limitation that study participants were preselected based on their risk profile with prevalent impaired glucose tolerance being a prerequisite. Consequently, diabetes incidence – even with lifestyle intervention – was very high, and thus, these results may not be easily generalizable to other risk groups.

60.8.2 Drugs in the Prevention of Type 2 Diabetes

Several trials have evaluated the efficacy of drugs in the prevention or delay of type 2 diabetes (Table 60.8). Metformin treatment resulted in a diabetes risk reduction in the diabetes prevention program (Knowler et al. 2002) and in the Indian Diabetes Prevention Program (Ramachandran et al. 2006). Metformin improves hyperglycemia primarily by suppressing hepatic gluconeogenesis. In the Study to Prevent Non-Insulin-Dependent Diabetes Mellitus (STOP-NIDDM) (Chiasson et al. 2002), the acarbose intervention group experienced a 25% diabetes risk reduction compared with the placebo group. Treatment with voglibose resulted in a risk reduction in a Japanese study (Kawamori et al. 2009). Both acarbose and voglibose are alpha-glucosidase inhibitors that decrease the absorption of carbohydrates from the intestine, resulting in a slower and lower rise in blood glucose, particularly after meals. Thiazolidinediones were used as drug intervention in several studies. Both troglitazone (Buchanan et al. 2002; Knowler et al. 2005) and rosiglitazone (Gerstein et al. 2006) reduced the risk of incident diabetes by at least 50% over 3 years compared to placebo. Furthermore, a low-dose combination therapy with rosiglitazone and metformin reduced the risk of type 2 diabetes in patients with impaired glucose tolerance (Zinman et al. 2010). The primary mechanism of action of thiazolidinediones involves binding to the peroxisome proliferator-activated receptor gamma, a transcription factor that regulates the expression and release of mediators of insulin resistance originating in adipose tissue. Treatment with orlistat, a potent inhibitor of pancreatic lipases that prevents the absorption of fats from the human diet, reduced diabetes risk by about 40% in two intervention studies (Heymsfield et al. 2000; Torgerson et al. 2004). A diabetes risk reduction was also observed for valsartan, an angiotensin receptor blocker (McMurray et al. 2010). In contrast, nateglinide, which stimulates insulin secretion, (Holman et al. 2010) and ramipril, an angiotensin-converting enzyme inhibitor (Bosch et al. 2006), were not found to significantly reduce diabetes risk in randomized controlled trials.

While these trials provide evidence that type 2 diabetes can be prevented through pharmacological interventions in high-risk populations, lifestyle modifications are more efficacious than drug interventions (Knowler et al. 2002; Ramachandran et al. 2006), and they do not cause side effects that are common in drug treatments. In particular, troglitazone was withdrawn from the market in 2000 due to liver toxicity, and rosiglitazone has been found to increase risk of cardiovascular disease. In addition, the beneficial effects of drugs on diabetes were lost shortly after the drugs were stopped (i.e., a few weeks), whereas lifestyle interventions had sustained benefits on diabetes prevention even after the active intervention discontinued for

Table 60.8 Major drug trials for prevention of type 2 diabetes

Study (author)	Population	Intervention	Relative risk (95% CI) of intervention compared with placebo
Diabetes Prevention Program (Knowler et al. 2002, 2005)	$n = 3{,}234$ with impaired glucose tolerance, aged ≥ 25 years	850 mg metformin or placebo twice daily, mean 2.8 years	0.69 (0.57–0.83)
		400 mg troglitazone or placebo daily, mean 0.9 years	0.25 ($p < 0.001$)
Indian Diabetes Prevention Program (Ramachandran et al. 2006)	$n = 269$ native Asian Indians with impaired glucose tolerance, aged 35–55 years	250 mg metformin or placebo twice daily, median 30 months	0.74 (0.65–0.81)
Study to Prevent Non-Insulin-Dependent Diabetes Mellitus (STOP-NIDDM) (Chiasson et al. 2002)	$n = 1{,}429$ with impaired glucose tolerance, aged 40–70 years	100 mg acarbose or placebo three times daily, mean 3.3 years	0.75 (0.63–0.90)
Japanese Study (Kawamori et al. 2009)	$n = 1{,}780$ with impaired glucose tolerance, aged 30–70 years	0.2 mg voglibose or placebo three times daily, mean 48 weeks	0.60 (0.43–0.82)
Three trials (Heymsfield et al. 2000)	$n = 675$ with BMI ≥ 30, aged > 18 years	120 mg orlistat or placebo three times daily, 582 days	0.39 ($p = 0.04$) among participants with impaired glucose tolerance at baseline
XENDOS (Torgerson et al. 2004)	$n = 3{,}305$ with BMI ≥ 30, aged 30–60 years	120 mg orlistat or placebo three times daily, 4 years	0.63 (0.46–0.86)
TRIPOD (Buchanan et al. 2002)	$n = 266$ hispanic women with previous gestational diabetes, aged ≥ 18 years	400 mg troglitazone or placebo once a day, mean 30 months	0.45 (0.25–0.83)
DREAM (Gerstein et al. 2006)	$n = 5{,}269$ with impaired fasting glucose or impaired glucose tolerance, aged ≥ 30 years	8 mg rosiglitazone or placebo daily	0.40 (0.35–0.46)
CANOE (Zinman et al. 2010)	$n = 207$ with impaired glucose tolerance, aged 30–75 years	2 mg rosiglitazone and 500 mg metformin or placebo twice daily, median 3.9 years	0.34 (0.20–0.59)
NAVIGATOR (Holman et al. 2010; McMurray et al. 2010)	$n = 9{,}306$ with impaired glucose tolerance, mean age 64 years	160 mg valsartan or placebo daily, median 5.0 years	0.86 (0.80–0.92)
		60 mg nateglinide or placebo three times daily, median 5.0 years	1.07 (1.00–1.15)
HOPE (Bosch et al. 2006)	$n = 5{,}269$ with impaired fasting glucose or impaired glucose tolerance, aged ≥ 30 years	15 mg ramipril or placebo daily, median 3 years	0.91 (0.81–1.03)

several years. Moreover, healthy diet and lifestyle modification is effective not only in preventing diabetes, but also in reducing the risk of other chronic diseases such as coronary heart disease (Ford et al. 2009; Stampfer et al. 2000).

Although lifestyle and drug interventions have been found to reduce the risk of diabetes among individuals at high risk, the benefits of these interventions for the prevention of cardiovascular events, death, or other long-term adverse health outcomes are yet to be demonstrated. For example, no difference in cardiovascular event rates was observed in the US Diabetes Prevention Program (Ratner et al. 2005); however, the study was not powered to examine cardiovascular outcomes. Treatment with acarbose reduced the risk of cardiovascular events in a post hoc analysis (Chiasson et al. 2003), but treatment groups had high attrition rates (33%). In the DREAM trial, cardiovascular event rates did not significantly differ between the rosiglitazone and placebo groups (Gerstein et al. 2006). Other studies among diabetic patients found that rosiglitazone treatment was associated with increased risk of myocardial infarction despite its glucose lowering effect (Nissen and Wolski 2007). On the other hand, lifestyle interventions, assuming a high adherence, are not only effective at reducing the risk of diabetes but also cost-effective at the population level. For a detailed discussion, the reader should refer to reviews on this topic (Waugh et al. 2007; Echouffo-Tcheugui et al. 2011; Norris et al. 2008).

60.9 Screening for Diabetes and Prediabetes

The benefits of early diabetes detection and treatment through screening programs compared to delayed treatment once cases become clinically diagnosed remains unclear because, to date, no randomized study has examined the effectiveness of screening programs. Results from the ADDITION study suggest that for screen-detected diabetes cases, intensive diabetes control did not significantly decrease the number of cardiovascular events beyond what has been achieved with standard care (Griffin et al. 2011). However, improvements in standard care over the study period may have resulted in small differences in patient care between the study groups. Notably, intensive glycemic control did not result in increased mortality risk – an effect that had been observed among patients with more advanced diabetes (Gerstein et al. 2008). As previously mentioned, thus far no randomized studies have compared screened and unscreened individuals with regard to long-term health benefits. The ongoing ADDITION-Cambridge study (Echouffo-Tcheugui et al. 2009) may shed more light on this question.

Despite uncertainties regarding the effectiveness of screening programs, a large body of literature has been published on screening techniques for undiagnosed diabetes. For an overview of the literature, the reader should refer to reviews on this topic (Echouffo-Tcheugui et al. 2009; Waugh et al. 2007). Point-of-care testing for FPG or HbA_{1c} is simple and convenient, but – as was discussed earlier in this chapter – they are unlikely to detect all prevalent diabetes cases. The 2hPG value is considered more reliable for diagnosing diabetes, but it requires an 8-h fast, commitment from the patient and the nursing staff, and has lower test-retest reproducibility compared with other tests. There is also a growing body of literature on risk scores that do not require blood sampling (Echouffo-Tcheugui et al. 2011). Although none of these tools are optimal in detecting undiagnosed diabetes, they have acceptable validity and are particularly useful tools to screen populations for further glucose screening.

Screening for diabetes cases at baseline and during follow-up has also been common practice in many analytical epidemiological studies. However, as pointed out earlier, screening at a single time point can produce a substantial number of false-positive screens, particularly when 2hPG is used. Because the diagnosis of diabetes requires confirmation in a subsequent assessment, screen-positive study participants should be considered to be potential cases but not necessarily true cases. While this approach allows to include undiagnosed cases, particularly the subgroup with isolated elevated 2hPG levels, the misclassification might in fact bias relative risk calculations and might lead to an overestimation of absolute risk. In contrast, other studies used only self-reported clinical cases. This approach will miss undiagnosed cases which can only be detected through screening and who can be expected to be on average younger than clinical cases. Also, clinical cases will frequently exclude asymptomatic diabetes cases with isolated elevated 2hPG levels since this criterion is not commonly used in clinical practice. Thus, studies based only on self-reported diabetes are likely to underestimate absolute risk and may not represent specific subgroups of cases adequately. While both approaches have pros and cons, the prevalence of undiagnosed diabetes cases in studies based on verified self-reports will not bias estimates of relative risk if this misclassification is non-differential with regard to exposure status (Greenland and Lash 2008). In general, it is, therefore, more relevant to invest in increasing the specificity of the case definition (by verification) than its sensitivity (by additional screening). Still, if studies are also required to accurately reflect absolute risks and to represent all patient subgroups, screening based on all diagnostic parameters and confirmation of screen-detected cases would need to be included.

There is a growing interest in developing prediction models for incident diabetes. Glucose screening could be used as a method to detect individuals at high risk with intermediate states of abnormal glucose regulation that precede overt type 2 diabetes. Impaired fasting glucose (FPG 100 or 110–125 mg/dL/5.6 or 6.1–6.9 mmol/L), impaired glucose tolerance (2hPG 140–199 mg/dL/7.8–11.0 mmol/L), or elevated HbA_{1c} (5.7–6.4%) have been used widely in this context to define a "prediabetic" status (Table 60.1). A meta-analysis of prospective studies conducted in different populations estimated that the relative risk for diabetes compared to normoglycemic individuals was 6.35 in people with impaired glucose tolerance, 5.52 in people with isolated impaired glucose tolerance, 4.66 in people with impaired fasting glucose, 7.54 in people with isolated impaired fasting glucose, and 12.13 in people with both impaired fasting glucose and impaired glucose tolerance (Gerstein et al. 2007). Despite demonstrating a clear increase in risk, dichotomizing glucose values as either "normal" or "prediabetic" neglects the continuum of risk associated with higher values. For example, a steady increase in diabetes risk has been observed for FPG values well within the range considered normal (Schulze et al. 2010b; Tirosh et al. 2005). Although current guidelines still include categories of prediabetes (Table 60.1), it is well established that risk is continuous, extending below the lower limit of the range for prediabetes and becoming disproportionately greater at the upper end of the range (American Diabetes Association 2011).

As an alternative to individual measures of blood glucose or HbA$_{1c}$, prediction models involving several diabetes risk factors hold significant promise for identifying those at high risk of developing type 2 diabetes. While several, moderately accurate diabetes prediction models have been developed based on risk factor information, there is evidence to suggest that these models could be improved with the addition of biochemical markers, in particular those reflecting hyperglycemia (Schmidt et al. 2005; Schulze et al. 2009). However, adding more complex markers of glucose and insulin metabolism, novel diabetes risk markers (such as CRP or adiponectin), or genetic information does not seem to further improve diabetes prediction. For a detailed review of prediction models, the reader should refer to Buijsse et al. (2011).

60.10 Conclusions

Similar to cardiovascular disease, overweight and obesity, diet, and lifestyle are predominant risk factors for type 2 diabetes. Diabetes epidemiology, therefore, shares important characteristics with other chronic disease epidemiology fields. Methodological developments to reduce misclassification in assessing dietary and lifestyle risk factors are of central interest to all fields of chronic disease epidemiology. Meanwhile, it is important to minimize confounding through improved study design and statistical analysis. Existing prospective cohort studies have increasingly been used for diabetes epidemiology although they may have initially been started as cancer epidemiology studies, e.g., the Nurses' Health Study or the European Prospective Investigation into Cancer (EPIC) study. The case-cohort study design for the evaluation of biochemical and genetic markers has increasingly been used to investigate type 2 diabetes and cardiovascular endpoints simultaneously to improve cost-efficiency. Recent advances in molecular and genetic epidemiology are of relevance to diabetes epidemiology, as they are to other fields.

Diabetes epidemiology also faces specific challenges. The definition of endpoints is heterogeneous across studies due to different diagnostic parameters and study-specific follow-up procedures. Based on current methodology, the misclassification of endpoints is almost unavoidable in diabetes epidemiology. This issue has also substantial implications for the estimation of population-level diabetes prevalence and incidence rates.

Diabetes epidemiology has played an essential role in demonstrating the importance of risk factors and the potential for prevention. The role of overweight and obesity in the etiology of type 2 diabetes has been demonstrated by numerous observational studies, and this link has also been causally established by intervention studies that focus on weight reduction among high-risk individuals. Physical activity, specific components of the diet, smoking, and other lifestyle factors have also been implicated in the development of diabetes. Overall, the majority of type 2 diabetes cases in the general population is attributable to modifiable diet and lifestyle factors.

These findings have had important implications for public health initiatives. For example, the International Diabetes Federation prevention plan for type 2 diabetes is based on controlling modifiable lifestyle risk factors. The plan divides the population into two groups that should be targeted simultaneously: people at high risk of developing type 2 diabetes (high-risk approach of prevention) and the remaining population (population-wide approach of prevention) (Alberti et al. 2007). Similarly, European guidelines for the prevention of type 2 diabetes have highlighted lifestyle modifications as key elements, both in high-risk and population-wide prevention approaches (Paulweber et al. 2010).

Accurate tools used to quantify absolute risk are essential for establishing programs that target high-risk individuals. The field of diabetes epidemiology has developed a number of tools that frequently outperform similar prediction tools used in the fields of cardiovascular or cancer epidemiology. Therefore, identifying high-risk groups seems a feasible task for diabetes interventions. However, the changes required to reduce the risk of diabetes at the population level are unlikely to be achieved without major environmental changes to facilitate appropriate choices by individuals.

References

Agardh EE, Ahlbom A, Andersson T, Efendic S, Grill V, Hallqvist J, Norman A, Ostenson CG (2003) Work stress and low sense of coherence is associated with type 2 diabetes in middle-aged Swedish women. Diabetes Care 26:719–724

Alberti KG, Zimmet P, Shaw J (2007) International Diabetes Federation: a consensus on Type 2 diabetes prevention. Diabet Med 24:451–463

American Diabetes Association (2011) Diagnosis and classification of diabetes mellitus. Diabetes Care 34(Suppl 1):S62–S69

Aune D, Ursin G, Veierod MB (2009) Meat consumption and the risk of type 2 diabetes: a systematic review and meta-analysis of cohort studies. Diabetologia 52:2277–2287

Baliunas DO, Taylor BJ, Irving H, Roerecke M, Patra J, Mohapatra S, Rehm J (2009) Alcohol as a risk factor for type 2 diabetes: a systematic review and meta-analysis. Diabetes Care 32: 2123–2132

Barr RG, Nathan DM, Meigs JB, Singer DE (2002) Tests of glycemia for the diagnosis of type 2 diabetes mellitus. Ann Intern Med 137:263–272

Beulens JW, van Beers RM, Stolk RP, Schaafsma G, Hendriks HF (2006) The effect of moderate alcohol consumption on fat distribution and adipocytokines. Obesity 14(Silver Spring):60–66

Beulens JW, van Loon LJ, Kok FJ, Pelsers M, Bobbert T, Spranger J, Helander A, Hendriks HF (2007) The effect of moderate alcohol consumption on adiponectin oligomers and muscle oxidative capacity: a human intervention study. Diabetologia 50:1388–1392

Beulens JW, de Zoete EC, Kok FJ, Schaafsma G, Hendriks HF (2008a) Effect of moderate alcohol consumption on adipokines and insulin sensitivity in lean and overweight men: a diet intervention study. Eur J Clin Nutr 62:1098–1105

Beulens JW, Rimm EB, Hu FB, Hendriks HF, Mukamal KJ (2008b) Alcohol consumption, mediating biomarkers, and risk of type 2 diabetes among middle-aged women. Diabetes Care 31:2050–2055

Bosch J, Yusuf S, Gerstein HC, Pogue J, Sheridan P, Dagenais G, Diaz R, Avezum A, Lanas F, Probstfield J, Fodor G, Holman RR (2006) Effect of ramipril on the incidence of diabetes. N Engl J Med 355:1551–1562

Brito EC, Lyssenko V, Renstrom F, Berglund G, Nilsson PM, Groop L, Franks PW (2009) Previously associated type 2 diabetes variants may interact with physical activity to modify the risk of impaired glucose regulation and type 2 diabetes: a study of 16,003 Swedish adults. Diabetes 58:1411–1418

Brohall G, Behre CJ, Hulthe J, Wikstrand J, Fagerberg B (2006) Prevalence of diabetes and impaired glucose tolerance in 64-year-old Swedish women: experiences of using repeated oral glucose tolerance tests. Diabetes Care 29:363–367

Browning LM, Hsieh SD, Ashwell M (2010) A systematic review of waist-to-height ratio as a screening tool for the prediction of cardiovascular disease and diabetes: 0.5 could be a suitable global boundary value. Nutr Res Rev 23:247–269

Buchanan TA, Xiang AH, Peters RK, Kjos SL, Marroquin A, Goico J, Ochoa C, Tan S, Berkowitz K, Hodis HN, Azen SP (2002) Preservation of pancreatic beta-cell function and prevention of type 2 diabetes by pharmacological treatment of insulin resistance in high-risk hispanic women. Diabetes 51:2796–2803

Buijsse B, Simmons RK, Griffin SJ, Schulze MB (2011) Risk assessment tools for identifying individuals at risk of developing type 2 diabetes. Epidemiol Rev 33:46–62

Butterfield WJ, Keen H, Whichelow MJ (1967) Renal glucose threshold variations with age. Br Med J 4:505–507

Cappuccio FP, D’Elia L, Strazzullo P, Miller MA (2010) Quantity and quality of sleep and incidence of type 2 diabetes: a systematic review and meta-analysis. Diabetes Care 33: 414–420

Carlsson S, Hammar N, Grill V (2005) Alcohol consumption and type 2 diabetes meta-analysis of epidemiological studies indicates a U-shaped relationship. Diabetologia 48:1051–1054

Carnethon MR, Sternfeld B, Schreiner PJ, Jacobs DR, Jr, Lewis CE, Liu K, Sidney S (2009) Association of 20-year changes in cardiorespiratory fitness with incident type 2 diabetes: the coronary artery risk development in young adults (CARDIA) fitness study. Diabetes Care 32:1284–1288

Carter P, Gray LJ, Troughton J, Khunti K, Davies MJ (2010) Fruit and vegetable intake and incidence of type 2 diabetes mellitus: systematic review and meta-analysis. BMJ 341:c4229

Chiasson JL, Josse RG, Gomis R, Hanefeld M, Karasik A, Laakso M (2002) Acarbose for prevention of type 2 diabetes mellitus: the STOP-NIDDM randomised trial. Lancet 359: 2072–2077

Chiasson JL, Josse RG, Gomis R, Hanefeld M, Karasik A, Laakso M (2003) Acarbose treatment and the risk of cardiovascular disease and hypertension in patients with impaired glucose tolerance: the STOP-NIDDM trial. JAMA 290:486–494

Colagiuri R, Short R, Buckley A (2010) The status of national diabetes programmes: a global survey of IDF member associations. Diabetes Res Clin Pract 87:137–142

Colagiuri S, Lee CM, Wong TY, Balkau B, Shaw JE, Borch-Johnsen K (2011) Glycemic thresholds for diabetes-specific retinopathy: implications for diagnostic criteria for diabetes. Diabetes Care 34:145–150

Collins VR, Dowse GK, Toelupe PM, Imo TT, Aloaina FL, Spark RA, Zimmet PZ (1994) Increasing prevalence of NIDDM in the Pacific island population of Western Samoa over a 13-year period. Diabetes Care 17:288–296

Cornelis MC, Qi L, Kraft P, Hu FB (2009a) TCF7L2, dietary carbohydrate, and risk of type 2 diabetes in US women. Am J Clin Nutr 89:1256–1262

Cornelis MC, Qi L, Zhang C, Kraft P, Manson J, Cai T, Hunter DJ, Hu FB (2009b) Joint effects of common genetic variants on the risk for type 2 diabetes in U.S. men and women of European ancestry. Ann Intern Med 150:541–550

Cowie CC, Rust KF, Byrd-Holt DD, Gregg EW, Ford ES, Geiss LS, Bainbridge KE, Fradkin JE (2010) Prevalence of diabetes and high risk for diabetes using A1C criteria in the U.S. population in 1988–2006. Diabetes Care 33:562–568

Davies MJ, Baer DJ, Judd JT, Brown ED, Campbell WS, Taylor PR (2002) Effects of moderate alcohol intake on fasting insulin and glucose concentrations and insulin sensitivity in postmenopausal women: a randomized controlled trial. JAMA 287:2559–2562

de Miguel-Yanes JM, Shrader P, Pencina MJ, Fox CS, Manning AK, Grant RW, Dupuis J, Florez JC, D'Agostino RB, Sr, Cupples LA, Meigs JB (2011) Genetic risk reclassification for type 2 diabetes by age below or above 50 years using 40 type 2 diabetes risk single nucleotide polymorphisms. Diabetes Care 34:121–125

de Munter JS, Hu FB, Spiegelman D, Franz M, van Dam RM (2007) Whole grain, bran, and germ intake and risk of type 2 diabetes: a prospective cohort study and systematic review. PLoS Med 4:e261

Ding EL, Hu FB (2007) Smoking and type 2 diabetes: underrecognized risks and disease burden. JAMA 298:2675–2676

Ding EL, Song Y, Malik VS, Liu S (2006) Sex differences of endogenous sex hormones and risk of type 2 diabetes: a systematic review and meta-analysis. JAMA 295:1288–1299

Diverse Populations Collaborative Group (2005) Weight-height relationships and body mass index: some observations from the Diverse Populations Collaboration. Am J Phys Anthropol 128: 220–229

Dong JY, Xun P, He K, Qin LQ (2011a) Magnesium intake and risk of type 2 diabetes: meta-analysis of prospective cohort studies. Diabetes Care 34:2116–2122

Dong JY, Zhang L, Zhang YH, Qin LQ (2011b) Dietary glycaemic index and glycaemic load in relation to the risk of type 2 diabetes: a meta-analysis of prospective cohort studies. Br J Nutr. doi:10.1017/S000711451100540X:1-6

Echouffo-Tcheugui JB, Simmons RK, Williams KM, Barling RS, Prevost AT, Kinmonth AL, Wareham NJ, Griffin SJ (2009) The ADDITION-Cambridge trial protocol: a cluster-randomised controlled trial of screening for type 2 diabetes and intensive treatment for screen-detected patients. BMC Public Health 9:136

Echouffo-Tcheugui JB, Ali MK, Griffin SJ, Narayan KM (2011) Screening for type 2 diabetes and dysglycemia. Epidemiol Rev 33:63–87

Ekoé JM, Rewers M, Williams R, Zimmet P (eds) (2008) The epidemiology of diabetes mellitus: An international perspective, 2nd edn. Wiley, Chichester

Eschwege E, Charles MA, Simon D, Thibult N, Balkau B (2001) Reproducibility of the diagnosis of diabetes over a 30-month follow-up: the Paris Prospective Study. Diabetes Care 24: 1941–1944

Esposito K, Kastorini CM, Panagiotakos DB, Giugliano D (2010) Prevention of type 2 diabetes by dietary patterns: a systematic review of prospective studies and meta-analysis. Metab Syndr Relat Disord 8:471–476

Fagot-Campagna A, Narayan KM, Hanson RL, Imperatore G, Howard BV, Nelson RG, Pettitt DJ, Knowler WC (1997) Plasma lipoproteins and incidence of non-insulin-dependent diabetes mellitus in Pima Indians: protective effect of HDL cholesterol in women. Atherosclerosis 128:113–119

Festa A, D'Agostino R Jr, Tracy RP, Haffner SM (2002) Elevated levels of acute-phase proteins and plasminogen activator inhibitor-1 predict the development of type 2 diabetes: the insulin resistance atherosclerosis study. Diabetes 51:1131–1137

Filozof C, Fernandez Pinilla MC, Fernandez-Cruz A (2004) Smoking cessation and weight gain. Obes Rev 5:95–103

Fisher E, Boeing H, Fritsche A, Doering F, Joost HG, Schulze MB (2009) Whole-grain consumption and transcription factor-7-like 2 (TCF7L2) rs7903146: gene-diet interaction in modulating type 2 diabetes risk. Br J Nutr 101:478–481

Florez JC (2008) Newly identified loci highlight beta cell dysfunction as a key cause of type 2 diabetes: where are the insulin resistance genes? Diabetologia 51:1100–1110

Florez JC, Jablonski KA, Kahn SE, Franks PW, Dabelea D, Hamman RF, Knowler WC, Nathan DM, Altshuler D (2007) Type 2 diabetes-associated missense polymorphisms KCNJ11 E23K and ABCC8 A1369S influence progression to diabetes and response to interventions in the Diabetes Prevention Program. Diabetes 56:531–536

Ford ES, Bergmann MM, Kroger J, Schienkiewitz A, Weikert C, Boeing H (2009) Healthy living is the best revenge: findings from the European Prospective Investigation Into Cancer and Nutrition-Potsdam study. Arch Intern Med 169:1355–1362

Forouhi NG, Harding AH, Allison M, Sandhu MS, Welch A, Luben R, Bingham S, Khaw KT, Wareham NJ (2007) Elevated serum ferritin levels predict new-onset type 2 diabetes: results from the EPIC-Norfolk prospective study. Diabetologia 50:949–956

Franks PW, Mesa JL, Harding AH, Wareham NJ (2007) Gene-lifestyle interaction on risk of type 2 diabetes. Nutr Metab Cardiovasc Dis 17:104–124s

Fraser A, Harris R, Sattar N, Ebrahim S, Davey Smith G, Lawlor DA (2009) Alanine aminotransferase, gamma-glutamyltransferase, and incident diabetes: the British Women's Heart and Health Study and meta-analysis. Diabetes Care 32:741–750

Fung TT, Schulze M, Manson JE, Willett WC, Hu FB (2004) Dietary patterns, meat intake, and the risk of type 2 diabetes in women. Arch Intern Med 164:2235–2240

Geiss LS, Pan L, Cadwell B, Gregg EW, Benjamin SM, Engelgau MM (2006) Changes in incidence of diabetes in U.S. adults, 1997–2003. Am J Prev Med 30:371–377

Gerstein HC, Yusuf S, Bosch J, Pogue J, Sheridan P, Dinccag N, Hanefeld M, Hoogwerf B, Laakso M, Mohan V, Shaw J, Zinman B, Holman RR (2006) Effect of rosiglitazone on the frequency of diabetes in patients with impaired glucose tolerance or impaired fasting glucose: a randomised controlled trial. Lancet 368:1096–1105

Gerstein HC, Santaguida P, Raina P, Morrison KM, Balion C, Hunt D, Yazdi H, Booker L (2007) Annual incidence and relative risk of diabetes in people with various categories of dysglycemia: a systematic overview and meta-analysis of prospective studies. Diabetes Res Clin Pract 78:305–312

Gerstein HC, Miller ME, Byington RP, Goff DC, Jr., Bigger JT, Buse JB, Cushman WC, Genuth S, Ismail-Beigi F, Grimm RH, Jr, Probstfield JL, Simons-Morton DG, Friedewald WT (2008) Effects of intensive glucose lowering in type 2 diabetes. N Engl J Med 358:2545–2559

Gill JM, Cooper AR (2008) Physical activity and prevention of type 2 diabetes mellitus. Sports Med 38:807–824

Gohdes D, Kaufman S, Valway S (1993) Diabetes in American Indians. An overview. Diabetes Care 16:239–243

Goldberg RB (2009) Cytokine and cytokine-like inflammation markers, endothelial dysfunction, and imbalanced coagulation in development of diabetes and its complications. J Clin Endocrinol Metab 94:3171–3182

Gopinath B, Rochtchina E, Flood VM, Mitchell P (2010) Healthy living and risk of major chronic diseases in an older population. Arch Intern Med 170:208–209

Greenland S, Lash TL (2008) Bias analysis. In: Rothman KJ, Greenland S and Lash TL (eds) Modern epidemiology, 3rd edn. LWW, Philadelphia, pp 345–380

Gregg EW, Cadwell BL, Cheng YJ, Cowie CC, Williams DE, Geiss L, Engelgau MM, Vinicor F (2004) Trends in the prevalence and ratio of diagnosed to undiagnosed diabetes according to obesity levels in the U.S. Diabetes Care 27:2806–2812

Griffin SJ, Borch-Johnsen K, Davies MJ, Khunti K, Rutten GE, Sandbæk A, Sharp SJ, Simmons RK, van den Donk M, Wareham NJ, Lauritzen T (2011) Effect of early intensive multifactorial therapy on 5-year cardiovascular outcomes in individuals with type 2 diabetes detected by screening (ADDITION-Europe): a cluster-randomised trial. Lancet 378:156–167

Grøntved A, Hu FB (2011) Television viewing and risk of type 2 diabetes, cardiovascular disease, and all-cause mortality: a meta-analysis. JAMA 305:2448–2455

Haffner SJ, Cassells H (2003) Hyperglycemia as a cardiovascular risk factor. Am J Med 115 (Suppl 8A):6S-11S

Halton TL, Willett WC, Liu S, Manson JE, Stampfer MJ, Hu FB (2006) Potato and french fry consumption and risk of type 2 diabetes in women. Am J Clin Nutr 83:284–290

Harding AH, Wareham NJ, Bingham SA, Khaw K, Luben R, Welch A, Forouhi NG (2008) Plasma vitamin C level, fruit and vegetable consumption, and the risk of new-onset type 2 diabetes mellitus: the European prospective investigation of cancer–Norfolk prospective study. Arch Intern Med 168:1493–1499

Heidemann C, Hoffmann K, Spranger J, Klipstein-Grobusch K, Mohlig M, Pfeiffer AF, Boeing H (2005) A dietary pattern protective against type 2 diabetes in the European Prospective Investigation into Cancer and Nutrition (EPIC) – Potsdam Study cohort. Diabetologia 48: 1126–1134

Heraclides A, Chandola T, Witte DR, Brunner EJ (2009) Psychosocial stress at work doubles the risk of type 2 diabetes in middle-aged women: evidence from the Whitehall II study. Diabetes Care 32:2230–2235

Herder C, Karakas M, Koenig W (2011) Biomarkers for the prediction of type 2 diabetes and cardiovascular disease. Clin Pharmacol Ther 90:52–66

Heymsfield SB, Segal KR, Hauptman J, Lucas CP, Boldrin MN, Rissanen A, Wilding JP, Sjostrom L (2000) Effects of weight loss with orlistat on glucose tolerance and progression to type 2 diabetes in obese adults. Arch Intern Med 160:1321–1326

Heymsfield SB, Heo M, Pietrobelli A (2011) Are adult body circumferences associated with height? Relevance to normative ranges and circumferential indexes. Am J Clin Nutr 93: 302–307

Hivert MF, Jablonski KA, Perreault L, Saxena R, McAteer JB, Franks PW, Hamman RF, Kahn SE, Haffner S, Meigs JB, Altshuler D, Knowler WC, Florez JC (2011) Updated genetic score based on 34 confirmed type 2 diabetes Loci is associated with diabetes incidence and regression to normoglycemia in the diabetes prevention program. Diabetes 60:1340–1348

Hodge AM, Dowse GK, Toelupe P, Collins VR, Imo T, Zimmet PZ (1994) Dramatic increase in the prevalence of obesity in western Samoa over the 13 year period 1978–1991. Int J Obes Relat Metab Disord 18:419–428

Holman RR, Haffner SM, McMurray JJ, Bethel MA, Holzhauer B, Hua TA, Belenkov Y, Boolell M, Buse JB, Buckley BM, Chacra AR, Chiang FT, Charbonnel B, Chow CC, Davies MJ, Deedwania P, Diem P, Einhorn D, Fonseca V, Fulcher GR, Gaciong Z, Gaztambide S, Giles T, Horton E, Ilkova H, Jenssen T, Kahn SE, Krum H, Laakso M, Leiter LA, Levitt NS, Mareev V, Martinez F, Masson C, Mazzone T, Meaney E, Nesto R, Pan C, Prager R, Raptis SA, Rutten GE, Sandstroem H, Schaper F, Scheen A, Schmitz O, Sinay I, Soska V, Stender S, Tamas G, Tognoni G, Tuomilehto J, Villamil AS, Vozar J, Califf RM (2010) Effect of nateglinide on the incidence of diabetes and cardiovascular events. N Engl J Med 362: 1463–1476

Howard BV, Manson JE, Stefanick ML, Beresford SA, Frank G, Jones B, Rodabough RJ, Snetselaar L, Thomson C, Tinker L, Vitolins M, Prentice R (2006) Low-fat dietary pattern and weight change over 7 years: the Women's Health Initiative Dietary Modification Trial. JAMA 295:39–49

Hu FB (2002) The impact of diabetes and prediabetes on risk of cardiovascular disease and mortality. Drugs Today (Barc) 38:769–775

Hu FB (2008) Obesity epidemiology. Oxford University Press, New York

Hu FB, Stampfer MJ, Rimm E, Ascherio A, Rosner BA, Spiegelman D, Willett WC (1999) Dietary fat and coronary heart disease: a comparison of approaches for adjusting for total energy intake and modeling repeated dietary measurements. Am J Epidemiol 149:531–540

Hu FB, Manson JE, Stampfer MJ, Colditz G, Liu S, Solomon CG, Willett WC (2001a) Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. N Engl J Med 345:790–797

Hu FB, van Dam RM, Liu S (2001b) Diet and risk of Type II diabetes: the role of types of fat and carbohydrate. Diabetologia 44:805–817

Huxley R, Lee CM, Barzi F, Timmermeister L, Czernichow S, Perkovic V, Grobbee DE, Batty D, Woodward M (2009) Coffee, decaffeinated coffee, and tea consumption in relation to incident type 2 diabetes mellitus: a systematic review with meta-analysis. Arch Intern Med 169: 2053–2063

Imamura F, Lichtenstein AH, Dallal GE, Meigs JB, Jacques PF (2009) Generalizability of dietary patterns associated with incidence of type 2 diabetes mellitus. Am J Clin Nutr 90:1075–1083

International Diabetes Federation (2009) Diabetes atlas. International Diabetes Federation, Brussels

Ix JH, Wassel CL, Kanaya AM, Vittinghoff E, Johnson KC, Koster A, Cauley JA, Harris TB, Cummings SR, Shlipak MG (2008) Fetuin-A and incident diabetes mellitus in older persons. JAMA 300:182–188

Jain SH, Massaro JM, Hoffmann U, Rosito GA, Vasan RS, Raji A, O'Donnell CJ, Meigs JB, Fox CS (2009) Cross-sectional associations between abdominal and thoracic adipose tissue compartments and adiponectin and resistin in the Framingham Heart Study. Diabetes Care 32:903–908

Jehn ML, Guallar E, Clark JM, Couper D, Duncan BB, Ballantyne CM, Hoogeveen RC, Harris ZL, Pankow JS (2007) A prospective study of plasma ferritin level and incident diabetes: the Atherosclerosis Risk in Communities (ARIC) Study. Am J Epidemiol 165:1047–1054

Jenkins DJ, Wolever TM, Taylor RH, Barker H, Fielden H, Baldwin JM, Bowling AC, Newman HC, Jenkins AL, Goff DV (1981) Glycemic index of foods: a physiological basis for carbohydrate exchange. Am J Clin Nutr 34:362–366

Jeon CY, Lokken RP, Hu FB, van Dam RM (2007) Physical activity of moderate intensity and risk of type 2 diabetes: a systematic review. Diabetes Care 30:744–752

Joosten MM, Beulens JW, Kersten S, Hendriks HF (2008) Moderate alcohol consumption increases insulin sensitivity and ADIPOQ expression in postmenopausal women: a randomised, crossover trial. Diabetologia 51:1375–1381

Kahn SE, Hull RL, Utzschneider KM (2006) Mechanisms linking obesity to insulin resistance and type 2 diabetes. Nature 444:840–846

Kanaya AM, Wassel Fyr C, Vittinghoff E, Harris TB, Park SW, Goodpaster BH, Tylavsky F, Cummings SR (2006) Adipocytokines and incident diabetes mellitus in older adults: the independent effect of plasminogen activator inhibitor 1. Arch Intern Med 166:350–356

Kastorini CM, Panagiotakos DB (2009) Dietary patterns and prevention of type 2 diabetes: from research to clinical practice; a systematic review. Curr Diabetes Rev 5:221–227

Kawamori R, Tajima N, Iwamoto Y, Kashiwagi A, Shimamoto K, Kaku K (2009) Voglibose for prevention of type 2 diabetes mellitus: a randomised, double-blind trial in Japanese individuals with impaired glucose tolerance. Lancet 373:1607–1614

Kendall CW, Josse AR, Esfahani A, Jenkins DJ (2010) Nuts, metabolic syndrome and diabetes. Br J Nutr 104:465–473

Khaw KT, Wareham N, Bingham S, Luben R, Welch A, Day N (2004) Association of hemoglobin A1c with cardiovascular disease and mortality in adults: the European prospective investigation into cancer in Norfolk. Ann Intern Med 141:413–420

Knol MJ, Twisk JW, Beekman AT, Heine RJ, Snoek FJ, Pouwer F (2006) Depression as a risk factor for the onset of type 2 diabetes mellitus. A meta-analysis. Diabetologia 49:837–845

Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, Nathan DM (2002) Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med 346:393–403

Knowler WC, Hamman RF, Edelstein SL, Barrett-Connor E, Ehrmann DA, Walker EA, Fowler SE, Nathan DM, Kahn SE (2005) Prevention of type 2 diabetes with troglitazone in the Diabetes Prevention Program. Diabetes 54:1150–1156

Knowler WC, Fowler SE, Hamman RF, Christophi CA, Hoffman HJ, Brenneman AT, Brown-Friday JO, Goldberg R, Venditti E, Nathan DM (2009) 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. Lancet 374:1677–1686

Koppes LL, Dekker JM, Hendriks HF, Bouter LM, Heine RJ (2005) Moderate alcohol consumption lowers the risk of type 2 diabetes: a meta-analysis of prospective observational studies. Diabetes Care 28:719–725

Kosaka K, Noda M, Kuzuya T (2005) Prevention of type 2 diabetes by lifestyle intervention: a Japanese trial in IGT males. Diabetes Res Clin Pract 67:152–162

Larsson SC, Wolk A (2007) Magnesium intake and risk of type 2 diabetes: a meta-analysis. J Intern Med 262:208–214

Lee CC, Adler AI, Sandhu MS, Sharp SJ, Forouhi NG, Erqou S, Luben R, Bingham S, Khaw KT, Wareham NJ (2009a) Association of C-reactive protein with type 2 diabetes: prospective analysis and meta-analysis. Diabetologia 52:1040–1047

Lee DC, Sui X, Church TS, Lee IM, Blair SN (2009b) Associations of cardiorespiratory fitness and obesity with risks of impaired fasting glucose and type 2 diabetes in men. Diabetes Care 32:257–262

Levitan EB, Song Y, Ford ES, Liu S (2004) Is nondiabetic hyperglycemia a risk factor for cardiovascular disease? A meta-analysis of prospective studies. Arch Intern Med 164: 2147–2155

Li G, Zhang P, Wang J, Gregg EW, Yang W, Gong Q, Li H, Jiang Y, An Y, Shuai Y, Zhang B, Zhang J, Thompson TJ, Gerzoff RB, Roglic G, Hu Y, Bennett PH (2008) The long-term effect of lifestyle interventions to prevent diabetes in the China Da Qing Diabetes Prevention Study: a 20-year follow-up study. Lancet 371:1783–1789

Li S, Shin HJ, Ding EL, van Dam RM (2009) Adiponectin levels and risk of type 2 diabetes: a systematic review and meta-analysis. JAMA 302:179–188

Lichtenstein AH, Schwab US (2000) Relationship of dietary fat to glucose metabolism. Atherosclerosis 150:227–243

Liese AD, Weis KE, Schulz M, Tooze JA (2009) Food intake patterns associated with incident type 2 diabetes: the Insulin Resistance Atherosclerosis Study. Diabetes Care 32:263–268

Lindstrom J, Ilanne-Parikka P, Peltonen M, Aunola S, Eriksson JG, Hemio K, Hamalainen H, Harkonen P, Keinanen-Kiukaanniemi S, Laakso M, Louheranta A, Mannelin M, Paturi M, Sundvall J, Valle TT, Uusitupa M, Tuomilehto J (2006) Sustained reduction in the incidence of type 2 diabetes by lifestyle intervention: follow-up of the Finnish Diabetes Prevention Study. Lancet 368:1673–1679

Liu S, Chou EL (2010) Dietary glycemic load and type 2 diabetes: modeling the glucose-raising potential of carbohydrates for prevention. Am J Clin Nutr 92:675–677

Lyssenko V, Jonsson A, Almgren P, Pulizzi N, Isomaa B, Tuomi T, Berglund G, Altshuler D, Nilsson P, Groop L (2008) Clinical risk factors, DNA variants, and the development of type 2 diabetes. N Engl J Med 359:2220–2232

Maahs DM, West NA, Lawrence JM, Mayer-Davis EJ (2010) Epidemiology of type 1 diabetes. Endocrinol Metab Clin North Am 39:481–497

Maarbjerg SJ, Sylow L, Richter EA (2011) Current understanding of increased insulin sensitivity after exercise – emerging candidates. Acta Physiol 202:323–335

Malik VS, Popkin BM, Bray GA, Despres JP, Willett WC, Hu FB (2010) Sugar-sweetened beverages and risk of metabolic syndrome and type 2 diabetes: a meta-analysis. Diabetes Care 33:2477–2483

Manson JE, Spelsberg A (1994) Primary prevention of non-insulin-dependent diabetes mellitus. Am J Prev Med 10:172–184

McCance DR, Hanson RL, Charles MA, Jacobsson LT, Pettitt DJ, Bennett PH, Knowler WC (1994) Comparison of tests for glycated haemoglobin and fasting and two hour plasma glucose concentrations as diagnostic methods for diabetes. BMJ 308:1323–1328

McMurray JJ, Holman RR, Haffner SM, Bethel MA, Holzhauer B, Hua TA, Belenkov Y, Boolell M, Buse JB, Buckley BM, Chacra AR, Chiang FT, Charbonnel B, Chow CC, Davies MJ, Deedwania P, Diem P, Einhorn D, Fonseca V, Fulcher GR, Gaciong Z, Gaztambide S, Giles T, Horton E, Ilkova H, Jenssen T, Kahn SE, Krum H, Laakso M, Leiter LA, Levitt NS, Mareev V, Martinez F, Masson C, Mazzone T, Meaney E, Nesto R, Pan C, Prager R, Raptis SA, Rutten GE, Sandstroem H, Schaper F, Scheen A, Schmitz O, Sinay I, Soska V, Stender S, Tamas G, Tognoni G, Tuomilehto J, Villamil AS, Vozar J, Califf RM (2010) Effect of valsartan on the incidence of diabetes and cardiovascular events. N Engl J Med 362:1477–1490

McNaughton SA, Mishra GD, Brunner EJ (2008) Dietary patterns, insulin resistance, and incidence of type 2 diabetes in the Whitehall II Study. Diabetes Care 31:1343–1348

Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PW, D'Agostino RB, Sr., Cupples LA (2008) Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med 359:2208–2219

Meisinger C, Strassburger K, Heier M, Thorand B, Baumeister SE, Giani G, Rathmann W (2010) Prevalence of undiagnosed diabetes and impaired glucose regulation in 35–59-year-old individuals in Southern Germany: the KORA F4 Study. Diabet Med 27:360–362

Mekary RA, Willett WC, Hu FB, Ding EL (2009) Isotemporal substitution paradigm for physical activity epidemiology and weight change. Am J Epidemiol 170:519–527

Melanson EL, Astrup A, Donahoo WT (2009) The relationship between dietary fat and fatty acid intake and body weight, diabetes, and the metabolic syndrome. Ann Nutr Metab 55:229–243

Mezuk B, Eaton WW, Albrecht S, Golden SH (2008) Depression and type 2 diabetes over the lifespan: a meta-analysis. Diabetes Care 31:2383–2390

Micha R, Wallace SK, Mozaffarian D (2010) Red and processed meat consumption and risk of incident coronary heart disease, stroke, and diabetes mellitus: a systematic review and meta-analysis. Circulation 121:2271–2283

Michels K, Schulze MB (2005) Can dietary patterns help us detect diet–disease associations? Nutr Res Rev 18:241–248

Montonen J, Drogan D, Joost HG, Boeing H, Fritsche A, Schleicher E, Schulze MB, Pischon T (2011) Estimation of the contribution of biomarkers of different metabolic pathways to risk of type 2 diabetes. Eur J Epidemiol 26:29–38

Mozaffarian D, Kamineni A, Carnethon M, Djousse L, Mukamal KJ, Siscovick D (2009) Lifestyle risk factors and new-onset diabetes mellitus in older adults: the cardiovascular health study. Arch Intern Med 169:798–807

Nissen SE, Wolski K (2007) Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N Engl J Med 356:2457–2471

Norris SL, Kansagara D, Bougatsos C, Fu R (2008) Screening adults for type 2 diabetes: a review of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med 148:855–868

Pan XR, Li GW, Hu YH, Wang JX, Yang WY, An ZX, Hu ZX, Lin J, Xiao JZ, Cao HB, Liu PA, Jiang XG, Jiang YY, Wang JP, Zheng H, Zhang H, Bennett PH, Howard BV (1997) Effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance. The Da Qing IGT and Diabetes Study. Diabetes Care 20:537–544

Pan A, Sun Q, Bernstein AM, Schulze MB, Manson JE, Willett WC, Hu FB (2011) Red meat consumption and risk of type 2 diabetes: 3 cohorts of US adults and an updated meta-analysis. Am J Clin Nutr 94:1088–1096

Paulweber B, Valensi P, Lindstrom J, Lalic NM, Greaves CJ, McKee M, Kissimova-Skarbek K, Liatis S, Cosson E, Szendroedi J, Sheppard KE, Charlesworth K, Felton AM, Hall M, Rissanen A, Tuomilehto J, Schwarz PE, Roden M, Paulweber M, Stadlmayr A, Kedenko L, Katsilambros N, Makrilakis K, Kamenov Z, Evans P, Gilis-Januszewska A, Lalic K, Jotic A, Djordevic P, Dimitrijevic-Sreckovic V, Huhmer U, Kulzer B, Puhl S, Lee-Barkey YH, AlKerwi A, Abraham C, Hardeman W, Acosta T, Adler M, Barengo N, Barengo R, Boavida JM, Christov V, Claussen B, Cos X, Deceukelier S, Djordjevic P, Fischer M, Gabriel-Sanchez R, Goldfracht M, Gomez JL, Handke U, Hauner H, Herbst J, Hermanns N, Herrebrugh L, Huber C, Huttunen J, Karadeniz S, Khalangot M, Kohler D, Kopp V, Kronsbein P, Kyne-Grzebalski D, Lalic N, Landgraf R, McIntosh C, Mesquita AC, Misina D, Muylle F, Neumann A, Paiva AC, Pajunen P, Peltonen M, Perrenoud L, Pfeiffer A, Polonen A, Raposo F, Reinehr T, Robinson C, Rothe U, Saaristo T, Scholl J, Spiers S, Stemper T, Stratmann B, Szybinski Z, Tankova T, Telle-Hjellset V, Terry G, Tolks D, Toti F, Undeutsch A, Valadas C, Velickiene D, Vermunt P, Weiss R, Wens J, Yilmaz T (2010) A European evidence-based guideline for the prevention of type 2 diabetes. Horm Metab Res 42(Suppl 1):S3–S36

Pouwer F, Kupper N, Adriaanse MC (2010) Does emotional stress cause type 2 diabetes mellitus? A review from the European Depression in Diabetes (EDID) Research Consortium. Discov Med 9:112–118

Priebe MG, van Binsbergen JJ, de Vos R, Vonk RJ (2008) Whole grain foods for the prevention of type 2 diabetes mellitus. Cochrane Database Syst Rev:CD006061

Qi L, Liang J (2010) Interactions between genetic factors that predict diabetes and dietary factors that ultimately impact on risk of diabetes. Curr Opin Lipidol 21:31–37

Qi L, Hu FB, Hu G (2008) Genes, environment, and interactions in prevention of type 2 diabetes: a focus on physical activity and lifestyle changes. Curr Mol Med 8:519–532

Qi L, Cornelis MC, Zhang C, van Dam RM, Hu FB (2009) Genetic predisposition, Western dietary pattern, and the risk of type 2 diabetes in men. Am J Clin Nutr 89:1453–1458

Qiu W, Rosner B (2010) Measurement error correction for the cumulative average model in the survival analysis of nutritional data: application to Nurses' Health Study. Lifetime Data Anal 16:136–153

Rajpathak SN, Crandall JP, Wylie-Rosett J, Kabat GC, Rohan TE, Hu FB (2009a) The role of iron in type 2 diabetes in humans. Biochim Biophys Acta 1790:671–681

Rajpathak SN, Wylie-Rosett J, Gunter MJ, Negassa A, Kabat GC, Rohan TE, Crandall J (2009b) Biomarkers of body iron stores and risk of developing type 2 diabetes. Diabetes Obes Metab 11:472–479

Ramachandran A, Snehalatha C, Mary S, Mukesh B, Bhaskar AD, Vijay V (2006) The Indian Diabetes Prevention Programme shows that lifestyle modification and metformin prevent type 2 diabetes in Asian Indian subjects with impaired glucose tolerance (IDPP-1). Diabetologia 49:289–297

Rathmann W, Haastert B, Icks A, Lowel H, Meisinger C, Holle R, Giani G (2003) High prevalence of undiagnosed diabetes mellitus in Southern Germany: target populations for efficient screening. The KORA survey 2000. Diabetologia 46:182–189

Ratner R, Goldberg R, Haffner S, Marcovina S, Orchard T, Fowler S, Temprosa M (2005) Impact of intensive lifestyle and metformin therapy on cardiovascular disease risk factors in the diabetes prevention program. Diabetes Care 28:888–894

Reis JP, Loria CM, Sorlie PD, Park Y, Hollenbeck A, Schatzkin A (2011) Lifestyle factors and risk for new-onset diabetes: a population-based cohort study. Ann Intern Med 155:292–299

Riserus U, Willett WC, Hu FB (2009) Dietary fats and prevention of type 2 diabetes. Prog Lipid Res 48:44–51

Salas-Salvado J, Bullo M, Babio N, Martinez-Gonzalez MA, Ibarrola-Jurado N, Basora J, Estruch R, Covas MI, Corella D, Aros F, Ruiz-Gutierrez V, Ros E (2011) Reduction in the incidence of type 2 diabetes with the Mediterranean diet: results of the PREDIMED-Reus nutrition intervention randomized trial. Diabetes Care 34:14–19

Sarwar N, Aspelund T, Eiriksdottir G, Gobin R, Seshasai SR, Forouhi NG, Sigurdsson G, Danesh J, Gudnason V (2010) Markers of dysglycaemia and risk of coronary heart disease in people without diabetes: Reykjavik prospective study and systematic review. PLoS Med 7:e1000278

Sattar N, Wannamethee SG, Forouhi NG (2008) Novel biochemical risk factors for type 2 diabetes: pathogenic insights or prediction possibilities? Diabetologia 51:926–940

Schmidt MI, Duncan BB, Bang H, Pankow JS, Ballantyne CM, Golden SH, Folsom AR, Chambless LE (2005) Identifying individuals at high risk for diabetes: the Atherosclerosis Risk in Communities study. Diabetes Care 28:2013–2018

Schulze MB, Hoffmann K (2006) Methodological approaches to study dietary patterns in relation to risk of coronary heart disease and stroke. Br J Nutr 95:860–869

Schulze MB, Hu FB (2002) Dietary patterns and risk of hypertension, type 2 diabetes mellitus, and coronary heart disease. Curr Atheroscler Rep 4:462–467

Schulze MB, Hu FB (2005) Primary prevention of diabetes: what can be done and how much can be prevented? Annu Rev Public Health 26:445–467

Schulze MB, Hoffmann K, Manson JE, Willett WC, Meigs JB, Weikert C, Heidemann C, Colditz GA, Hu FB (2005) Dietary pattern, inflammation, and incidence of type 2 diabetes in women. Am J Clin Nutr 82:675–684; quiz 714–675

Schulze MB, Heidemann C, Schienkiewitz A, Bergmann MM, Hoffmann K, Boeing H (2006) Comparison of anthropometric characteristics in predicting the incidence of type 2 diabetes in the EPIC-Potsdam Study. Diabetes Care 29:1921–1923

Schulze MB, Schulz M, Heidemann C, Schienkiewitz A, Hoffmann K, Boeing H (2007) Fiber and magnesium intake and incidence of type 2 diabetes: a prospective study and meta-analysis. Arch Intern Med 167:956–965

Schulze MB, Schulz M, Heidemann C, Schienkiewitz A, Hoffmann K, Boeing H (2008) Carbohydrate intake and incidence of type 2 diabetes in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam Study. Br J Nutr 99:1107–1116

Schulze MB, Weikert C, Pischon T, Bergmann MM, Al-Hasani H, Schleicher E, Fritsche A, Haring HU, Boeing H, Joost HG (2009) Use of multiple metabolic and genetic markers to improve the prediction of type 2 diabetes: the EPIC-Potsdam Study. Diabetes Care 32: 2116–2119

Schulze MB, Rathmann W, Giani G, Joost HG (2010a) Diabetesprävalenz: Verlässliche Schätzungen stehen noch aus. Dtsch Arztebl 107:A 1694–1696 (only in German)

Schulze MB, Fritsche A, Boeing H, Joost HG (2010b) Fasting plasma glucose and type 2 diabetes risk: a non-linear relationship. Diabet Med 27:473–476

Sierksma A, Patel H, Ouchi N, Kihara S, Funahashi T, Heine RJ, Grobbee DE, Kluft C, Hendriks HF (2004) Effect of moderate alcohol consumption on adiponectin, tumor necrosis factor-alpha, and insulin sensitivity. Diabetes Care 27:184–189

Siperstein MD (1975) The glucose tolerance test: a pitfall in the diagnosis of diabetes mellitus. Adv Intern Med 20:297–323

Sluijs I, Beulens JW, van der A DL, Spijkerman AM, Grobbee DE, van der Schouw YT (2010) Dietary intake of total, animal, and vegetable protein and risk of type 2 diabetes in the European Prospective Investigation into Cancer and Nutrition (EPIC)-NL study. Diabetes Care 33:43–48

Song Y, Manson JE, Buring JE, Liu S (2004) A prospective study of red meat consumption and type 2 diabetes in middle-aged and elderly women: the women's health study. Diabetes Care 27:2108–2115

Song Y, Cook NR, Albert CM, Van Denburgh M, Manson JE (2009) Effects of vitamins C and E and beta-carotene on the risk of type 2 diabetes in women at high risk of cardiovascular disease: a randomized controlled trial. Am J Clin Nutr 90:429–437

Souverein OW, Dekkers AL, Geelen A, Haubrock J, de Vries JH, Ocke MC, Harttig U, Boeing H, van 't Veer P (2011) Comparing four methods to estimate usual intake distributions. Eur J Clin Nutr 65 (Suppl 1):S92–101

Stampfer MJ, Hu FB, Manson JE, Rimm EB, Willett WC (2000) Primary prevention of coronary heart disease in women through diet and lifestyle. N Engl J Med 343:16–22

Stefan N, Fritsche A, Weikert C, Boeing H, Joost HG, Haring HU, Schulze MB (2008) Plasma fetuin-A levels and the risk of type 2 diabetes. Diabetes 57:2762–2767

Stern MP, Williams K, Haffner SM (2002) Identification of persons at high risk for type 2 diabetes mellitus: do we need the oral glucose tolerance test? Ann Intern Med 136:575–581

Stranges S, Marshall JR, Natarajan R, Donahue RP, Trevisan M, Combs GF, Cappuccio FP, Ceriello A, Reid ME (2007) Effects of long-term selenium supplementation on the incidence of type 2 diabetes: a randomized trial. Ann Intern Med 147:217–223

Stumvoll M, Goldstein BJ, van Haeften TW (2005) Type 2 diabetes: principles of pathogenesis and therapy. Lancet 365:1333–1346

Talmud PJ, Hingorani AD, Cooper JA, Marmot MG, Brunner EJ, Kumari M, Kivimaki M, Humphries SE (2010) Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. BMJ 340:b4838

Taylor AE, Ebrahim S, Ben-Shlomo Y, Martin RM, Whincup PH, Yarnell JW, Wannamethee SG, Lawlor DA (2010) Comparison of the associations of body mass index and measures of central adiposity and fat mass with coronary heart disease, diabetes, and all-cause mortality: a study using data from 4 UK cohorts. Am J Clin Nutr 91:547–556

Tinker LF, Bonds DE, Margolis KL, Manson JE, Howard BV, Larson J, Perri MG, Beresford SA, Robinson JG, Rodriguez B, Safford MM, Wenger NK, Stevens VJ, Parker LM (2008) Low-fat dietary pattern and risk of treated diabetes mellitus in postmenopausal women: the Women's Health Initiative randomized controlled dietary modification trial. Arch Intern Med 168: 1500–1511

Tirosh A, Shai I, Tekes-Manova D, Israeli E, Pereg D, Shochat T, Kochba I, Rudich A (2005) Normal fasting plasma glucose levels and type 2 diabetes in young men. N Engl J Med 353:1454–1462

Tong X, Dong JY, Wu ZW, Li W, Qin LQ (2011) Dairy consumption and risk of type 2 diabetes mellitus: a meta-analysis of cohort studies. Eur J Clin Nutr 65:1027–1031

Torgerson JS, Hauptman J, Boldrin MN, Sjostrom L (2004) XENical in the prevention of diabetes in obese subjects (XENDOS) study: a randomized study of orlistat as an adjunct to lifestyle changes for the prevention of type 2 diabetes in obese patients. Diabetes Care 27:155–161

Tuomilehto J, Lindstrom J, Eriksson JG, Valle TT, Hamalainen H, Ilanne-Parikka P, Keinanen-Kiukaanniemi S, Laakso M, Louheranta A, Rastas M, Salminen V, Uusitupa M (2001) Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. N Engl J Med 344:1343–1350

van Dam RM, Hu FB (2005) Coffee consumption and risk of type 2 diabetes: a systematic review. JAMA 294:97–104

van Dam RM, Rimm EB, Willett WC, Stampfer MJ, Hu FB (2002) Dietary patterns and risk for type 2 diabetes mellitus in U.S. men. Ann Intern Med 136:201–209

van Hoek M, Dehghan A, Witteman JC, van Duijn CM, Uitterlinden AG, Oostra BA, Hofman A, Sijbrands EJ, Janssens AC (2008) Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. Diabetes 57:3122–3128

Vazquez G, Duval S, Jacobs DR, Jr., Silventoinen K (2007) Comparison of body mass index, waist circumference, and waist/hip ratio in predicting incident diabetes: a meta-analysis. Epidemiol Rev 29:115–128

Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, Wu G, Willer CJ, Raychaudhuri S, McCarroll SA, Langenberg C, Hofmann OM, Dupuis J, Qi L, Segre AV, van Hoek M, Navarro P, Ardlie K, Balkau B, Benediktsson R, Bennett AJ, Blagieva R, Boerwinkle E, Bonnycastle LL, Bengtsson Bostrom K, Bravenboer B, Bumpstead S, Burtt NP, Charpentier G, Chines PS, Cornelis M, Couper DJ, Crawford G, Doney AS, Elliott KS, Elliott AL, Erdos MR, Fox CS, Franklin CS, Ganser M, Gieger C, Grarup N, Green T, Griffin S, Groves CJ, Guiducci C, Hadjadj S, Hassanali N, Herder C, Isomaa B, Jackson AU, Johnson PR, Jorgensen T, Kao WH, Klopp N, Kong A, Kraft P, Kuusisto J, Lauritzen T, Li M, Lieverse A, Lindgren CM, Lyssenko V, Marre M, Meitinger T, Midthjell K, Morken MA, Narisu N, Nilsson P, Owen KR, Payne F, Perry JR, Petersen AK, Platou C, Proenca C, Prokopenko I, Rathmann W, Rayner NW, Robertson NR, Rocheleau G, Roden M, Sampson MJ, Saxena R, Shields BM, Shrader P, Sigurdsson G, Sparso T, Strassburger K, Stringham HM, Sun Q, Swift AJ, Thorand B, Tichet J, Tuomi T, van Dam RM, van Haeften TW, van Herpt T, van Vliet-Ostaptchouk JV, Walters GB, Weedon MN, Wijmenga C, Witteman J, Bergman RN, Cauchi S, Collins FS, Gloyn AL, Gyllensten U, Hansen T, Hide WA, Hitman GA, Hofman A, Hunter DJ, Hveem K, Laakso M, Mohlke KL, Morris AD, Palmer CN, Pramstaller PP, Rudan I, Sijbrands E, Stein LD, Tuomilehto J, Uitterlinden A, Walker M, Wareham NJ, Watanabe RM, Abecasis GR, Boehm BO, Campbell H, Daly MJ, Hattersley AT, Hu FB, Meigs JB, Pankow JS, Pedersen O, Wichmann HE, Barroso I, Florez JC, Frayling TM, Groop L, Sladek R, Thorsteinsdottir U, Wilson JF, Illig T, Froguel P, van Duijn CM, Stefansson K, Altshuler D, Boehnke M, McCarthy MI (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 42:579–589

Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, McCabe E, Lewis GD, Fox CS, Jacques PF, Fernandez C, O'Donnell CJ, Carr SA, Mootha VK, Florez JC, Souza A, Melander O, Clish CB, Gerszten RE (2011) Metabolite profiles and the risk of developing diabetes. Nat Med 17: 448–453

Wannamethee SG, Shaper AG, Perry IJ (2001) Smoking as a modifiable risk factor for type 2 diabetes in middle-aged men. Diabetes Care 24:1590–1595

Wareham NJ (2007) Epidemiological studies of physical activity and diabetes risk, and implications for diabetes prevention. Appl Physiol Nutr Metab 32:778–782

Waugh N, Scotland G, McNamee P, Gillett M, Brennan A, Goyder E, Williams R, John A (2007) Screening for type 2 diabetes: literature review and economic modelling. Health Technol Assess 11:iii–iv, ix–xi, 1–125

Weitzman S, Wang CH, Pankow JS, Schmidt MI, Brancati FL (2010) Are measures of height and leg length related to incident diabetes mellitus? The ARIC (Atherosclerosis Risk in Communities) study. Acta Diabetol 47:237–242

Will JC, Galuska DA, Ford ES, Mokdad A, Calle EE (2001) Cigarette smoking and diabetes mellitus: evidence of a positive association from a large prospective cohort study. Int J Epidemiol 30:540–546

Willi C, Bodenmann P, Ghali WA, Faris PD, Cornuz J (2007) Active smoking and the risk of type 2 diabetes: a systematic review and meta-analysis. JAMA 298:2654–2664

Wilson PW, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB, Sr (2007) Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. Arch Intern Med 167:1068–1074

World Health Organization (2006) Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: report of a WHO/IDF consultation. World Health Organization, Geneva

World Health Organization (2011) Use of glycated haemoglobin (HbA1c) in the diagnosis of diabetes mellitus: abbreviated report of a WHO consultation. World Health Organization, Geneva

Yeh HC, Duncan BB, Schmidt MI, Wang NY, Brancati FL (2010) Smoking, smoking cessation, and risk for type 2 diabetes mellitus: a cohort study. Ann Intern Med 152:10–17

Zilkens RR, Burke V, Watts G, Beilin LJ, Puddey IB (2003) The effect of alcohol intake on insulin sensitivity in men: a randomized controlled trial. Diabetes Care 26:608–612

Zinman B, Harris SB, Neuman J, Gerstein HC, Retnakaran RR, Raboud J, Qi Y, Hanley AJ (2010) Low-dose combination therapy with rosiglitazone and metformin to prevent type 2 diabetes mellitus (CANOE trial): a double-blind randomised controlled study. Lancet 376:103–111

Index

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0,